

Too hesitant steps forward in European collaboration

Researchers seeking to build projects that make the best of Europe's potential have no obvious place to turn. A new proposal should be actively encouraged by funding agencies, despite their strong reservations.

Most people would think it self-evident that, reflecting the development of political and economic structures, there should be an organization that coordinates and possibly funds pan-European collaborative research projects. The European Commission already does so in many areas, but has set its face (and indeed is legally constrained) against the support of basic science for its own sake. The comparatively tiny European Science Foundation (ESF), among other functions, successfully coordinates research networks that are supported financially by Europe's national funding agencies — the same bodies that fund the ESF itself and which together control about 95% of Europe's research funds. The question that has only now reached the top of the European policy agenda is whether some of those funds should be funnelled towards collaboration in a more systematic and institutionalized fashion.

Two bodies that are dealing with this question are the ESF itself, for whom the idea represents an obvious opportunity, and an informal group that meets but twice a year: the heads of European research councils (Eurohorcs), for whom the obstacles to the idea loom large — possibly too large — while the more sceptical of them question the need for any new arrangements. Multilateral collaboration is abundant, after all. All too quickly, the sceptics say, any move forward will mushroom into their ultimate bogeyman conjured up to frighten overly Europhilic politicians: the European Research Council — a bureaucratic monster bound to erode national scientific accountability.

Catalyst for action

There are clearly topics of research — for example, in basic environmental and geoscientific research, in the development of medium-sized facilities, in phase I clinical trials, in social sciences — where it would be beneficial to have an organization known by researchers and agencies alike as the first point of contact and, where enough pan-European interest is apparent, as the catalyst of multilateral action. Politically, the only way forward to that end is a stepwise development that gingerly but demonstrably explores ways of delivering such a benefit without unleashing a self-perpetuating and insufficiently accountable bureaucracy.

A timely if not entirely original proposal has recently provided a possible way forward and usefully stimulated discussion among the *cognoscenti* of European science policy. Reinder van Duinen, head of the Netherlands Organization for Scientific Research (NWO), suggests that a body such as the ESF should coordinate research projects that, at the least, would bring together existing or nascent national projects in particular areas (see *Nature* 395, 208; 1998). In preliminary clinical trials or studies of genetic associations, for example, the benefits would flow from the harmonization of methodologies and the scale and diversity of the population samples studied. At the most, says van Duinen, the ESF might, in the exploratory phase, administer a pot of contributions from interested research councils for competitive bids from across Europe, subject to peer review, towards a particular scientific goal.

This latter suggestion is where the critical obstacles appear most distinctly. First, it seems unlikely that some research councils (France's CNRS and the UK research councils not least) could enter into such an arrangement without government permission, raising the unappetizing prospect of many negotiations over national accountability. Second, it could be seen as the birth of a vaunting new Eurobureaucracy, leading to the European Research Council. Third, it requires that the ESF (assuming that it would be undesirable to invent yet another new body) be trusted to do the job well, and in a manner that suits research councils and researchers.

This last point begs further critical questions. On paper the ESF is the natural organization for the job. Founded nearly 25 years ago, it has given excellent and welcomed advice to the European Commission on the troubled fifth Framework research programme, has acted as midwife to at least one international facility, and has helped European astronomers fight off some of the worst impacts of satellite-based mobile communications, while at least some of its research networks have been well endorsed by participants. But its leaders in recent years have failed to secure the confidence of its funders, ambiguously labelled "member organizations", that it has a vital role to play.

Persuasiveness

The ESF is now under the relatively new leadership of its Spanish director-general, Enric Banda. Of more fundamental importance, the relationship of the ESF to its member organizations, the European research councils, and also to research communities themselves, are under active reconsideration. There is therefore an opportunity for the ESF to seize van Duinen's proposal as a way forward, but to do so in a way that does not undermine its current activities. This will be a test of the combined persuasiveness of Banda and the ESF's long-serving and influential president, Dai Rees. It requires above all a demonstration by them that, perhaps with the judicious appointment of a heavyweight and diplomatically skilled senior manager responsible for the project, the ESF can gain the trust of the research agencies that their own interests will not be compromised.

The final question begged is how researchers themselves express their opinions and requirements amid all of this, both now, in the initial phases of the proposal, and in any structure that might emerge. Euroscience, a fledgling grass-roots body inspired by the model of the American Association for the Advancement of Science (AAAS), is being mentioned by some research council heads as potentially providing a vehicle. With luck, imagination and good judgement, Euroscience should play a useful role in Europe. But there is no more reason to wish to defer to it in the present context than there would be for the US National Institutes of Health to depend on the views of the AAAS.

Thus the bodies that most desire a streamlining and encouragement of European collaboration in basic science presently lack the required strength, while the group most capable of ensuring that something happens is only minimally enthusiastic. While that situation persists, scientists in Europe will remain inadequately served. □



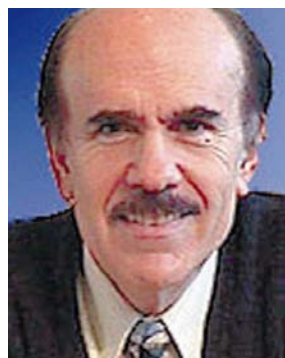
Nobel award stirs up debate on nitric oxide breakthrough

[LONDON] The Nobel committee has once again sparked controversy, particularly in Britain, with the award this week of the 1998 prize for physiology or medicine to pharmacologists Robert Furchgott, of the State University of New York, Louis Ignarro, of the University of California, Los Angeles, and Ferid Murad, of University of Texas Medical School in Houston.

No-one is disputing that the pioneering work of the three US researchers on nitric oxide as a signalling molecule in the cardiovascular system (see box) deserves the recognition of the committee. But many feel that a fourth name should also have been recognized — that of Salvador Moncada, currently director of the Wolfson Institute for Biomedical Research at University College London.

The controversy originates in 1987, when a paper by Ignarro confirming the identity of endothelium-derived relaxing factor (EDRF) and nitric oxide, as had been hypothesized separately by him and Furchgott, appeared in the *Proceedings of the National Academy of Sciences* (84, 9265–9269; 1987).

The publication came six months after a similar report in *Nature* by Moncada and colleagues, working in London, who had already concluded that EDRF and nitric oxide were identical (see *Nature* 327, 524–526; 1987). Writing in the same issue of *Nature*, Paul Vanhoutte, then at the Mayo Clinic in Rochester, Minnesota, acclaimed the finding as “the climax of one of the most exciting sagas in vascular physiology and



The winners: Robert Furchgott (left), Louis Ignarro and Ferid Murad.

pharmacology”. Moncada’s achievements were recognized when he was elected as a foreign associate of the National Academy of Sciences in 1994. The citation says that he “discovered that mammalian vascular tissues generate nitric oxide that is biosynthesized from L-arginine. He elucidated the relevance of this pathway as a universal transduction mechanism for the regulation of cell function and communication”.

Moncada expressed surprise on Monday (12 October) at the decision of the Nobel committee to award the medicine prize to Furchgott, Murad and Ignarro. His disappointment was shared by Sir John Vane, joint winner of the 1982 Nobel prize for his work on prostaglandins.

Although delighted that the committee has recognized the importance of the nitric oxide field, Vane expressed regret that Moncada’s contributions were not acknowl-

edged. But he believes that the prize is thoroughly deserved by Furchgott — without whom “the field would not even exist” — and says that he had nominated both Furchgott and Ignarro for the prize.

More outspoken support for Moncada comes from Nobel laureate Cesar Milstein of



Moncada: disappointed at lack of recognition.

the MRC Laboratory of Molecular Biology in Cambridge. Milstein won the 1984 Nobel prize for his work on monoclonal antibodies, and dug deep into the nitric oxide literature following the award in 1996 of the prestigious Albert Lasker Medical Research

Award to Furchgott and Murad.

Milstein argues on the basis of his investigations that Moncada was the first to take seriously the suggestion by Furchgott and Ignarro that EDRF might be identical with nitric oxide — and did the key experiments that made the earlier hypothesis physiologically meaningful.

Characterizing the decision as “scandalous”, Milstein argues that the committee made a serious blunder by excluding Moncada from the awards.

Milstein says he does not believe that the omission of Moncada will seriously damage British research. But he regrets that this research will not get the fillip it deserves, and admits that the omission leaves him “uneasy about the way that recognition is being given at the level of major prizes, and in terms of careers and publications”.

Not all researchers are as critical. Jonathan Stamler, of Duke University Medical Center in North Carolina, who also works on nitric oxide in the cardiovascular

Hidden role of a gas of many parts

[LONDON] Interest in nitric oxide has exploded in recent years due to the recognition that the gas plays an important role in many physiological processes, and that manipulating the nitric oxide signalling pathway can have major medical benefits. But, only 30 years ago, the idea that a gas could have biological functions would have seemed far-fetched.

In 1980, Robert Furchgott reported that acetylcholine could contract or relax the smooth muscle surrounding blood vessels, but that this relaxation only occurred if the endothelium — the lining

of the blood vessel — was intact (see *Nature* 288, 373–376; 1980). The implication was that a signal molecule was being released by endothelial cells and acting on the muscle cells, and Furchgott called the postulated molecule endothelium-derived relaxing factor — EDRF.

The link with nitric oxide was not immediately obvious. EDRF was highly labile, and candidate molecules, such as prostacyclin and other prostanoids, were soon ruled out. But in the 1970s Ferid Murad had discovered that

the gas could activate guanylate cyclase, the enzyme that mediates the production of the signalling molecule cyclic GMP.

Murad also showed that nitroglycerin and other vasodilating drugs released nitric oxide. But the physiological relevance of this discovery remained unknown. The various links in the story were only pulled together at a conference in 1986 when Furchgott and Louis Ignarro, who had been working independently of each other, hypothesized that EDRF and nitric oxide were identical.

R.H.

system, says he is very pleased that the work of Furchgott, Ignarro and Murad has been recognized.

Stamler accepts that Moncada "had the vision to appreciate the field in the broader sense and to shoulder the burden of later research". But he believes that earlier work had already opened the way for the later discoveries, as the Nobel committee has recognized.

Solomon Snyder of Johns Hopkins University in Baltimore, Maryland, says he agrees with this interpretation, and accepts that the committee would have done much scholarly work before making its decision.

He says the medical significance of Murad's work on nitroglycerin is enormous. Murad showed that the active metabolite of nitroglycerin is nitric oxide, leading drug companies to search for other drugs that could release the gas.

Ironically, Alfred Nobel, the inventor of dynamite — in which the explosiveness of nitroglycerin is kept in check by a porous material composed of diatoms — was ordered by his doctor to "eat" nitroglycerin for a heart condition, but refused to take it.

Nitroglycerin is a potent drug now used in the treatment of angina. Just 0.5 mg placed under the tongue can make a patient suffering an angina attack feel better within minutes, due to the dilation of the affected blood vessels.

It is now clear that nitric oxide is the major determinant of blood pressure. But its medical significance does not stop there. In 1992 Snyder and colleagues showed that nitric oxide synthase (NOS), the enzyme that produces nitric oxide, is expressed in neurons in the penis and that nitric oxide mediates erectile function.

Rory Howlett

Physicists rewarded for 'fractional electrons'



Physics laureates: Robert Laughlin (left) shortly after hearing of his award, Daniel Tsui (centre) and Horst Störmer.

[LONDON] This year's Nobel prize for physics has been awarded to the researchers who first observed and explained the fractional quantum Hall effect. This is the effect in which an electric current within a two-dimensional conducting material appears to be made up of charge carriers bearing a fraction of the charge on an electron.

Horst Störmer of Columbia University, New York, and Bell Laboratories, and Daniel Tsui of Princeton University, who both saw the effect experimentally in 1982, share the prize with Robert Laughlin of Stanford University, California, who provided a theoretical explanation shortly afterwards.

The standard Hall effect is the lateral deflection of moving charge carriers — an electric current — in a magnetic field. It was discovered in 1879 by Edwin Hall, and today provides the basis for determining the charge and density of charge carriers in a semiconductor (electrons and holes are deflected in different directions).

In 1980 Klaus von Klitzing found that, when the charge carriers are confined within a very thin conducting film (that is, in two dimensions), the magnitude of the Hall current (or, equivalently, the conductance of the material) no longer varies smoothly with magnetic-field strength at very low temperatures.

Instead, the conductance varies with field strength in a series of abrupt steps. In other words, the conductance is quantized: it changes in integral multiples of the fundamental quantum unit of conductance, e^2/h (where e is the charge on the electron and h is Planck's constant).

The fractional quantum Hall effect represents a deeper puzzle, since it seems to reveal a change in the nature of the fundamental particles. Much the same can be said of superconductivity, in which electrons appear to attract one another (or, more properly, to show bosonic instead of fermionic behaviour), and of superfluidity, in which the atoms of the superfluid no longer generate viscosity.

The FQHE was seen by Tsui and Störmer for the transport of a two-dimensional electron 'gas' in a semiconductor heterostructure fabricated by Art Gossard, now at the University of California at Santa Barbara. On applying magnetic fields of up to 30 tesla to a sample cooled to about a tenth of a degree kelvin, they observed jumps in conductance with a value of $e^2/3h$, implying that the charge carriers had a fractional charge of $e/3$. Subsequent studies revealed charges of $2e/5$, $3e/7$ and other (odd-denominator) fractions.

Laughlin proposed that the magnetic flux lines penetrating the sample encourage the charge carriers to condense into quasiparticles. He demonstrated that such quasiparticles act as though they have fractional charges with the values seen in the experiments.

The crucial insight, says Moty Heiblum of the Weizmann Institute in Israel, was the recognition of the role of electron correlations. In semiconductor physics, says Heiblum, "all of us managed to work with a single-electron picture for many years". But the study of strongly correlated electrons in solid-state physics has now become an important field of research.

Philip Ball

Theoretical chemistry makes its mark

[LONDON] The award of the Nobel prize for chemistry to Walter Kohn of the University of California at Santa Barbara and John Pople of Northwestern University, Illinois, signals a recognition that computational chemistry is now a tool at the chemist's disposal to equal any experimental or analytical technique.

Kohn provided the theoretical framework for calculating the electronic structure of molecules without having to grapple with the formidable task of solving the full Schrödinger equation. He showed that the total energy of a system can be expressed simply in terms of the distribution of its electron density, without regard to the details of the electron motions; the density is a function of the spatial coordinates, and the energy is a 'function of a function', or a functional.

Kohn developed what was later to be called density functional theory in 1964, with applications in physics in mind. This theory is now widely used to calculate the



Pople: work has 'led to an industry'.

electronic band structures of solids, and also in liquid-state physics.

The application of approximate computational methods to chemistry, meanwhile, was pioneered by Pople.

Towards the end of the 1960s he developed the GAUSSIAN-70 program

for calculating the electronic structure of molecules and the nature of their interactions and reactions.

Current modifications of this program are now used by thousands of chemists throughout the world, according to Roberto Car of the Ecole Polytechnique Fédérale in Lausanne, Switzerland. "[Pople] has built an industry around this," says Car.

That view is echoed by Nicholas Handy at the University of Cambridge, who feels that the laureates are "exactly the right two people" to be recognized.

P.B.

Congress warns nuclear labs of spy risk

[WASHINGTON] The three nuclear weapons laboratories of the US Department of Energy are tightening controls on access by foreign scientists in the wake of growing concern in Congress that the lack of thorough vetting of visitors is enabling foreign intelligence services to spy on them.

The directors of the Los Alamos, Sandia and Lawrence Livermore laboratories told a congressional hearing on 6 October that steps were already being taken to vet visitors more carefully, following a 1997 report from the General Accounting Office (GAO) that attributed security breaches to inadequate vetting.

But Duncan Hunter (Republican, California), chairman of the procurement subcommittee of the powerful National Security Committee in the House of Representatives, which held the hearing, doubts that the laboratories or the Department of Energy are doing enough to vet foreign visitors.

According to one laboratory official, Hunter and other committee members from both parties are "really steamed" about the security issue.

The weapons laboratories have opened up markedly to overseas visitors since the end of the Cold War. According to the GAO, the number of foreign visitors to the three laboratories has grown from around 3,800 in 1988 to 5,983 in 1994 and 6,998 in 1996. Almost one-third of visitors now come from countries that the United States regards as "sensitive" — chiefly India, China and Russia.

Foreign visitors include hundreds from Russian weapons laboratories engaged in 'lab-to-lab' collaborations on weapons issues, as well as some Chinese weapons scientists involved in a far smaller and less formal liaison with their US counterparts.

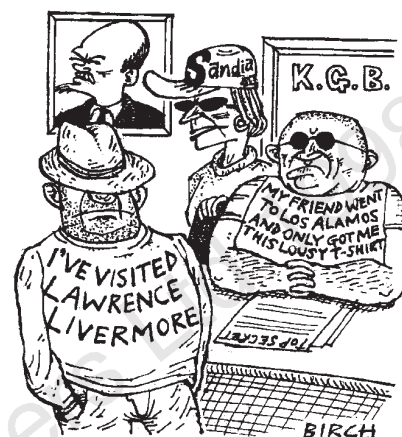
Scientists from all over the world also work with the laboratories on non-weapons-related work, which has expanded greatly since the end of the Cold War, and on topics such as inertial confinement fusion, which is unclassified but could still have weapons applications.

Additionally, the laboratories hire some young scientists who are not US citizens, especially in disciplines where US-born PhDs are in short supply.

Most of these visitors do not gain access to the closed areas of the laboratories where classified nuclear weapons work is done. But Congress is concerned about secret evidence that trained spies have been able to enter the non-classified areas of the laboratories, and perhaps extract valuable information.

According to a GAO report published in September last year, examples of espionage activities against the labs "include recent cases involving the possible theft or compromise of sensitive information in which foreign nationals at the Department of Energy's laboratories played a prominent role". Hunter's subcommittee was extensively briefed on these breaches during a closed session of last week's hearing.

The GAO said that Los Alamos and Sandia



were performing background checks in advance of visits of scientists from sensitive countries in only 5 per cent of cases; in contrast, Livermore performed them on almost half of such visitors. It found 13 cases in which "persons with suspected foreign intelligence connections" were allowed access to Los Alamos or Sandia between 1994 and 1996.

Both the Department of Energy and the laboratories say they are revamping their counter-intelligence procedures and will implement more background checks. At Sandia, for example, a 1997 security review has resulted in new requirements for advanced notice of visits. This may be up to 55 days for scientists from sensitive countries who plan to work for more than a month in a closed part of the laboratory.

"The rules are much, much stricter, and we feel we've got control," says Melanie Flores, head of counter-intelligence at the New Mexico laboratory.

"It adds a little bit of a burden, but foreign visitors tend to plan far in advance anyway," says John Crawford, Sandia's vice-president. The new rules do not deter visitors, Crawford says, but even tighter controls could do just that. "You could tighten things so far that people would be totally discouraged from visiting the labs — it definitely could be damaging," he says.

In testimony to the committee, John Browne, director of Los Alamos, stressed the importance of interchange with foreign scientists: "Without participation in the international scientific community, the credibility of and talent in these labs will wither."

But officials say that Congress is inclined to push the labs towards a more cautious approach to such participation.

The GAO is already working with more than one congressional committee to investigate the subject further. If Congress is not satisfied with the answers, its options include reducing funding for collaborative programmes, and enforcing stricter rules that may keep scientists from sensitive countries out of the laboratories.

Colin Macilwain

Pay crisis drives Russian scientists abroad

[MOSCOW] The flight of scientists from Russia shows no signs of slowing down. The Ministry of Science and Technology released figures last week showing that 15,000 scientists have left the country during the past five to seven years.

The emigrants represent five per cent of Russia's 300,000 scientists. Most went to the United States, with a minority heading for Israel, escaping continued political instability, a crisis over unpaid salaries, and the falling value of the ruble.

The average monthly salary of a scientist in Russia is 1,000 rubles (US\$60). Although this is a 60 per cent increase from 1997, this gain has been wiped out by the devaluation of the ruble in August to 17 rubles per US dollar. Before devaluation, one US dollar was worth 6.35 rubles.

Despite repeated promises, the prospects that salaries will be paid on time remain poor. Initial hopes that the new prime minister, Yevgeny Primakov, himself a fellow of the Russian Academy of Sciences, would resolve the crisis have quickly faded.

Although Primakov has promised to find a solution, he is expected to give greater priority to the plight of miners and teachers, who face a similar crisis.

Primakov has allowed the science and technology portfolio — including virtually all research policy — to remain with Vladimir Bulgak, the deputy prime minister, who was science minister under the previous government.

But Bulgak's ability to lobby for scientists will be limited. He is in overall charge of 18 government agencies and departments including the ministries of transport, fuel and energy, and atomic energy.

In a recent speech, Yuri Luzhkov, the politically ambitious mayor of Moscow, called on Bulgak to "return the lost valuables" to Russian science if impending catastrophe was to be avoided. Luzhkov said that 15 to 20 years are needed for Russia's science to reach the level of developed countries — and then only if financing reaches "adequate" levels.

Carl Levitin

Japan may require labels on genetic food

[TOKYO] Japan has announced plans to introduce tough regulations that would require all food containing genetically modified organisms (GMOs) to be labelled as such, despite strong opposition from the domestic food industry and overseas food importers. Following a nationwide opinion poll that closed on 9 October, the government is expected to decide shortly on which of two alternative plans to adopt.

Under one version of its proposal, products that may contain protein and genetic material of GMOs would have to be labelled as such. This suggests that products such as maize and soybeans, which are mostly imported from the United States, as well as processed food containing such ingredients, would have to be tested.

The US government has already sent Trade Representative Charlene Barshefsky to Tokyo to discuss the issue. At a meeting last month with Shoichi Nakagawa, the Japanese agriculture minister, Barshefsky warned that

the regulations could jeopardize trade relations between Japan and the United States.

Because it produces only 30 per cent of its own agricultural needs, Japan is highly dependent on imports, especially from the United States, which provides 80 per cent of Japan's total soybean supply and most of the 20 genetically modified products currently available in Japan.

The proposed regulations were announced by the Ministry of Agriculture, Forestry and Fisheries (MAFF) in August. They have generated widespread surprise, because an apparent deadlock between consumer groups and the food industry on GMO labelling had made it difficult for the ministry to reach a consensus.

Those most startled by the agriculture ministry's announcement were, ironically, members of its own committee on GMO food labelling, who had been discussing the issue since May last year.

Yasue Ito, a committee member and direc-



Telling signs: food would have to be labelled if it contained herbicide-resistant soya (right).

tor of the Consumer Science Federation, says the committee's discussions went only as far as whether the presence or absence of GMOs should be the basis of labelling.

"MAFF was meant to come up with a preliminary plan based on issues discussed by the committee, but instead we were presented with an outline of regulations involving issues we have never discussed," says Ito.

Strong lobbying from the food industry against GMO labelling had suggested that the ministry would agree that labels need only be used to indicate the non-use of GMOs. But MAFF has proposed two alternative plans, both of which involve labelling all food products containing GMOs, although in one case this would be mandatory and in the other it would be voluntary.

One plan stipulates that if the crop can be identified at an early stage as "genetically modified" or "non-genetically modified", the products resulting from such crops will be labelled accordingly; products resulting from the mixture of GMOs and non-GMOs would be labelled as "undifferentiated".

Meanwhile, representatives of the Japanese food industry are concerned about the difficulty of tracing original ingredients in processed food, as well as the potentially high cost of carrying out tests to identify products containing the DNA and proteins of GMOs. They also point out that there are no established techniques in Japan for testing for the presence of molecules from GMOs.

"It is technically possible to detect genetic material and protein of GMOs using a method based on the polymerase chain reaction," says Akihiro Hino from MAFF's National Food Research Institute. "The real problems are those associated with policy-making — issues such as setting a lower limit for the presence of molecules from GMOs and dealing with the problem of contamination."

The ministry's decision on which version of its proposal to adopt is to be announced next month, based on the poll's results. "Public opinion will be the crucial factor in reaching our final decision, as the whole purpose of GMO labelling is to recognize the consumers' wish to know what they are eating," says a MAFF spokesman.

Asako Saegusa

UK environment report looks to the people

[LONDON] The public should be more closely involved in setting local and national environmental standards if these are to be widely accepted, according to Britain's Royal Commission on Environmental Pollution.

In a report published last week, the commission is reluctant to make specific recommendations about how this should be achieved. But it stresses the general need for a "new approach to policy-making".

The commission is chaired by Sir Tom Blundell, head of the department of biochemistry at the University of Cambridge and former chief executive of the Biotechnology and Biological Sciences Research Council. Its report, *Setting Environmental Standards*, argues that both scientific evidence and social values need to be used in setting environmental standards. In the past, it says, the neglect of values and non-scientific opinions has resulted in poor public acceptance of environmental policies.

The report stresses that scientific assessments of environmental issues remain an important basis for decision-makers. But it says that such assessments should present a range of possibilities for action, rather than a single option.

"Scientists should not usurp the policy-maker's role," says Clair Chilvers, an epidemiologist at the University of Nottingham and a member of the Royal Commission. Chilvers points out that in many cases, such as the disposal of the Brent Spar oil platform in the North Sea or the propagation of genetically modified crops, scientific assessment on its own failed to

provide a firm basis for policy decisions, partly because it contained many uncertainties.

Decision-makers should recognize that the "requirement for sound science as the basis for environmental policy is not the requirement for absolute knowledge", says the report. They should also accept that there are bound to be "limitations and uncertainties" at each stage of a scientific assessment of an environmental issue.

In addition to the traditional procedure of direct, science-based, government regulation of environmental issues, new instruments should be introduced. The commission says these could include negotiated agreements between governments and industries or companies, and economic instruments, such as green taxes or charges.

To ensure that people's values, attitudes and opinions are adequately taken into account, the commission recommends that new methods should be added to the "relatively technocratic procedures" of setting environmental standards. These might include community forums, citizens' juries and consensus conferences.

The commission recommends that the UK Department of the Environment should incorporate such methods into the procedures for considering environmental issues and setting standards. The use of new methods for determining public values should also be "high on the agenda" for European institutions, says the report.

Quirin Schiermeier

UK nuclear waste firm faces US critics

[WASHINGTON] A top official of BNFL, the UK-based nuclear-fuels corporation, told a congressional hearing last week that the company is ready to tackle the waste left behind from the plutonium production that sustained the US nuclear arsenal during the Cold War.

BNFL is confident it can vitrify ten per cent of the 54 million gallons of liquid, solid and sludge waste held in 177 underground tanks at Hanford in Washington state, the ugliest nuclear waste problem in the United States. "There's no question that this can be done," says Tom Crimmins, BNFL's US chief executive officer.

But the contract proposed by the Department of Energy (DoE) with BNFL was criticized at a hearing of the oversight subcommittee of the House Commerce Committee for leaving too much risk with government. The contract specifies that BNFL will design and build waste treatment and vitrification plants at Hanford to begin operation by 2007.

Despite the DoE describing the contract as

a "privatization" of the Hanford waste problem, it was criticized by Joe Barton (Republican, Texas), chair of the subcommittee.

Barton said that BNFL would commit \$300 million to the project but receive \$1.8 billion in profits and fees. He promised that Congress will monitor the contract closely.

BNFL was left as the sole candidate after a rival contractor, Lockheed, was rejected by the DoE in May on technical grounds.

In August, the DoE signed a contract with BNFL to process ten per cent of the waste by 2017 for \$6.9 billion. This price will be reviewed in two years, by which time BNFL is due to have built a pilot \$25 million waste treatment facility and completed 30 per cent of the plant's design work.

BNFL will borrow capital from the private sector to build the vitrification plant, then DoE will pay it to vitrify the waste. Critics, including some environmentalists, say this is a false economy, as the DoE is liable for BNFL's debts if the contract is not completed.

A report from the General Accounting

Office last week said "DoE's financial risks are significant", and doubted that the department could adequately oversee the project.

But some critics go further. Gerald Pollet, director of the environmental group Heart of America Northwest, told the hearing the government could save \$3 billion by contracting on a conventional, fixed-price basis.

But Crimmins says BNFL will carry the risk of cost overruns. He is backed up by the energy undersecretary Ernie Moniz, who says: "There's plenty of risk for the contractor — their equity is first in line."

As well as the Hanford contract, BNFL is leading a consortium that hopes to take over the nuclear operations of Westinghouse, including management of the Savannah River site in South Carolina.

Environmental groups have criticized BNFL for its allegedly secretive culture and its track record on health and safety in Britain. But no one at the hearing objected to these or the fact that it is publicly owned by the British government.

Colin Macilwain

Epidemiologist's funds axed after report on Californian smoking

[SAN FRANCISCO] The state of California's decision to stop funding an epidemiologist, following a dispute over his report on smoking behaviour, has prompted concern about the state's apparent reluctance to accept unfavourable scientific results.

John Pierce, an epidemiologist at the University of California, San Diego, who has conducted three surveys on smoking for the California Department of Health Services during the past decade, has had his contract abruptly discontinued after a dispute over both his methodology and his findings.

Pierce had been sent a letter in December confirming the agency's intention to extend his contract up to the year 2000 for \$4.85 million. He was asked to assess adolescent smoking, complete a fourth survey on smoking behaviour in the state, and report on the success of a state tobacco control programme funded by a cigarette tax of 25 cents a pack introduced in 1990.

But after months of wrangling over Pierce's 1996 final report, the department accepted his results, but ended his contract.

The report, available on the World-Wide Web (ssdc.ucsd.edu/tobacco/reports/), concludes that the tobacco control programme lost its effectiveness, and a downward trend in smoking rates ended, in 1994, the year state legislature cut spending on anti-smoking campaigns by 40 per cent. It also predicts that smoking in young people is likely to have increased by 14 per cent between 1996 and 1999.

Donald Lyman, chief of the Division of



Smoke-filled rooms: California's crackdown on smoking has led to the rise of smoking clubs.

Chronic Disease and Injury Control, says that, in an environment already strained by an anxious state administration and a hostile tobacco industry, Pierce had become impossible to work with.

But Pierce and other researchers claim the department's action was intended to prevent the publication of data that highlight the harm caused to smoking-prevention efforts by state funding cuts.

"He produced results [that were] not politically acceptable and got fired for it," says Stanton Glantz, a tobacco researcher and professor of medicine at the University of California, San Francisco. He calls the decision "outrageous and appalling".

State health department officials say that, among other weaknesses, the analysis relies on questionable statistical models and incorporates a problematic definition of individuals who smoke less than 100 cigarettes a year.

But Pierce argues that the methodological adjustments he was asked to make would have altered the results and created a continuing downward trend in smoking rates. He did make some of the changes, but told the department that the others would damage the scientific credibility of the report.

"We felt we'd bent over backwards to be responsive to comments," says Pierce. He says he thinks the department was concerned about his report's clear documentation of the effect of funding cuts and so decided not to renew his contract.

Joel Moskowitz, deputy director of the Center for Family and Community Health at the University of California, Berkeley, says there is merit to the department's concerns. He criticizes several aspects of Pierce's results and analysis.

But Moskowitz says he has no doubt that the state axed Pierce for political reasons after he refused to make the proposed changes. Moskowitz suggests this reflects problems throughout the state, where the agency that is funded to carry out a programme is also asked to evaluate it.

But Lyman denies that the non-renewal of Pierce's contract was intended to penalize him for bringing bad news.

Sally Lehrmann

Canadians call for tax cuts to curb the 'brain drain'...

[MONTREAL] Several influential business groups are urging Canada's federal government to cut income taxes in an attempt to dissuade Canadian scientists and technologists from emigrating to greener pastures, notably the United States.

The groups argue that action is needed urgently. Nick Cercone, chairman of the University of Waterloo's computer sciences department, says about 100 of spring's 240 graduates would be taking jobs in the United States. It has also been reported that about half of this year's 80 computer-graphics graduates from Sheridan College near Toronto have already moved across the border.

The business groups are supporting the Liberal government's controversial policy of financing tax cuts from a billion-dollar surplus in the national unemployment benefit scheme (UI).

The UI has produced annual surpluses for the past five years that currently total about Can\$20 billion (US\$13 billion). Legislation requires that such surpluses are used to reduce premiums or increase benefits.

In recent years, however, the government has cut benefits and increased premiums, and has dipped into the surplus to reduce its overall budget deficit. It has also proposed changing the UI legislation to avoid having to reduce premiums or increase benefits.

This has provoked an unusual display of unity among opposition parties, which say they will resort to extreme tactics to block the Liberals' plans. They are proposing a commission made up of labour and business representatives to run the UI plan, and want the government to be prohibited by law from

using the funds for any other purpose.

But the Business Council on National Issues (BCNI), made up of the chief executive officers of 150 companies, said last week that UI premiums should be left basically intact, and surplus money should be used to help lower taxes next year.

According to Thomas d'Aquino, the group's president, the size of the UI premium cut sought by some other groups and political parties is so big—about Can\$7 billion—that it would wipe out the surplus next year.

In contrast, argues d'Aquino, cutting income tax would both encourage economic competition and help persuade Canada's best graduates to stay at home.

Support for the government also comes from Sharon Glover, vice-president of the Canadian Chamber of Commerce (COC), which represents 170,000 businesses. She agrees with the BCNI that tax cuts are a greater priority, and that her organization's members are increasingly worried about a 'brain drain' from Canada.

The extent of the exodus is expected to be revealed in a publication to be produced soon by the C. D. Howe Institute, a think tank.

Finance minister Paul Martin said recently that, although tax reduction is a government priority, it would not completely stem the flow of skilled workers out of the country.

Canada has to compete not only with higher pre-tax salary packages in the United States, but also with greater opportunities for premiums and stock options. Once Canadian tax levels are reduced, the government will turn its attention to these other issues, said Martin.

David Spurgeon

... as skilled workers head for the United States

[MONTREAL] Canada is increasingly becoming the target of recruitment drives by US companies. In 1995, representatives of 30 US companies visited the University of Waterloo, but this number rose to 100 the following year and 170 in 1997.

Canada's Software Human Resources Council estimates a shortfall of about 20,000 information technology specialists, partly as a result of the emigration of graduates.

Since May 1997, the government has admitted more than 830 such

specialists from foreign countries under a pilot programme co-sponsored by the council. The programme speeds up applications for work visas for specific categories of worker deemed in short supply.

The specialists, from countries including the United States, France, India, China, Singapore and Russia, earn average salaries of Can\$49,500 (US\$32,000). Another government pilot project recently announced also allows the spouses of such workers to find jobs.

The situation is slightly

different for medical staff. The *Canadian Medical Association Journal* reports that the number of physicians moving abroad fell from 731 in 1996 to 659 in 1997, the lowest level since 1993.

Of those moving, 60 per cent were aged 40 or younger and male. Of the specialists, 63 per cent worked in clinical medicine or laboratory fields. Taking into account physicians returning from abroad, the net loss fell from 513 in 1996 to 452 in 1997, still more than twice the 1990 net loss of just 207 physicians.

D.S.

France abandons plan to reform national biomedical agency

[PARIS] A controversial proposed reform of the French national biomedical agency, INSERM, has been quietly shelved. The government is reported to be drafting a new plan that takes greater account of the general need to improve the coordination of life sciences research in France.

The reform plan was intended to modify the statutes of the agency and streamline its internal decision-making (see *Nature* 391, 110 & 392, 9; 1998). A final text of the reform decree was adopted by INSERM's board of directors in the spring after a series of concessions to trade unions.

The decree had been challenged by many scientists who claimed it was poorly thought through and likely to have little impact. It was also seen as threatening to damage INSERM's capacity to carry out fundamental research.

Françoise Cavaillé, the national secretary of the trade union of scientific researchers at INSERM, this week criticized what she described as the enormous time and effort that had been wasted by scientists at the agency in the long negotiations over the now aborted reform decree.

Cavaillé argues that much of this could have been avoided if the Ministry of National Education, Research and Technology had decided to carry out broader consultation on the reform before pressing forward unilaterally.

The unions complained that the decree would transfer considerable control over strategic decisions on research directions and funding from the agency's scientific bodies to boards made up largely of officials appointed by the ministry (see *Nature* 393, 506; 1998). The decree had already been emptied of its most controversial elements, such as the plan to split INSERM into five departments.

A spokesperson for INSERM says it is unclear what will happen next. But, according to one observer, the ministry now intends to review the structure of INSERM within a broader reform of the life sciences. Last month it set up a national coordinating committee for research in the life sciences, made up of representatives of the nine public research organizations with interests in the life sciences, and leading scientists (see *Nature* 395, 315; 1998).

The committee's task will be to propose and evaluate programmes coordinated between the different research agencies, and to advise the ministry on strategy. The ministry could not be reached for comment this week on its current thinking about INSERM's future.

Declan Butler

UK contract staff need help, says report

[LONDON] Principal investigators in the United Kingdom are being urged to do more to ensure that contract researchers working on projects under their responsibility receive proper guidance on career prospects. This might include the steps necessary to move into a non-academic career.

Similarly, bodies such as the research councils, which support contract researchers, and the higher-education institutions at which most of them work, have been encouraged to increase their own efforts to ensure that such individuals are aware of the support already on offer.

For example, it is recommended that every higher-education institution should establish a clear focus for contract research staff. And research councils have been asked to consider whether grant evaluation procedures might be adopted to reward effective management and career guidance by the institutions.

These are among the proposals contained in a report published last week by the Research Careers Initiative (RCI), which was set up in 1996 to oversee the implementation of a 'concordat' between representatives of higher-education institutions and research councils on the career management of contract researchers.

The signing of the concordat reflected a growing recognition that the issue is a major problem for the research community. It is calculated that there are currently 30,000 researchers in British universities working on contracts of an average of three years, compared with 70,000 full-time staff.

In some disciplines the figures are even more dramatic; indeed, in science and engineering there are around two contract researchers for every three full-time acad-

Main recommendations

The main recommendations of the Research Careers Initiative are that:

- A clear focus should be established on policy for contract research staff at every institution of higher education.
- Institutions should obtain

feedback on questionnaires completed by contract researchers about their working practices.

- A guide to best practice produced by the RCI should be widely publicized.
- Careers guidance resources should

be increased.

- A framework for the continuing personal and professional development of contract researchers should be developed.
- The work of the RCI should be continued by a successor body.

emic researchers. Many take on a succession of contract positions in the hope of eventually being appointed to an academic post; at present, only 20 per cent achieve this.

"It is very important to manage expectations," says David Clark of the Engineering and Physical Sciences Research Council (EPSRC). "Contract researchers should not expect that the next step is automatically an academic position."

The EPSRC has taken the lead in introducing measures to ensure that contract researchers are allowed to develop their careers in a way that will make them attractive to potential employers outside universities. For example, it has agreed to pay for contract researchers to go on certain career development courses.

Gareth Roberts, vice-chancellor of the University of Sheffield and chairman of the RCI, says in an introduction to the report that such initiatives and other moves by universities themselves illustrate that "measurable progress" has been made since the concordat was signed. But he admits that "problems persist".

The report and the surveys on which it is based describe some of the problems. For

example, it says that although the number of higher-education institutions taking action is growing, the proportion involved is still "far too low". It suggests that research funders should adopt "a more pro-active role".

Questionnaires distributed among contract researchers revealed that they usually receive little career guidance. Although this is partly because there is little such guidance, the RCI admits that it is also the result of a disinclination among many to seek it.

An RCI working group identified several reasons for the low participation in career guidance activities. These range from a lack of support among research managers, to concern that an expression of interest in an alternative career could destroy already slim prospects of an academic career.

A separate working group on training argues that there seems to be a weakness in the research management chain at the level of principal investigator. According to the working group, principal investigators are often under great pressure to achieve a high level of research output and to meet contractual commitments to research sponsors. "In these circumstances, the personal and development needs of contract research staff are not always high on the agenda."

The RCI's main report has been welcomed by the Association of University Teachers, the labour union representing the interests of university staff — including contract researchers. "Its recommendations offer hope of real progress in careers guidance and professional development for university researchers," says the union's general secretary David Triesman.

Triesman goes further, however, arguing against what he calls "the contract culture", where basic insecurity creates "the inefficiency of high turnover, wastage, non-completion and inadequate dissemination of research".

Although, in publishing the report, the RCI has achieved the objective it was set up for, one recommendation is that it be reconstituted as a smaller group to help take forward some of the proposals in its report.

The full text of the report can be found on: www.cvc.ac.uk

David Dickson

New telescope shows off its paces

[LONDON] This image of the Dumbbell Nebula (right) has been obtained by the European Southern Observatory's Very Large Telescope on Mount Paranal in Chile.

The Dumbbell Nebula is located in the Vulpecula constellation, believed to be around 1,200 light years from Earth. It consists of very rarified gas that has been ejected from a hot central star, visible in the image, and which is now in one of the final stages of its evolution.

The nebula's gas atoms are heated by the intense ultraviolet radiation from the star, and emit radiation at specific wavelengths. Filters on the telescope allow the isolation of emissions from particular atoms and ions, shown in the intricate structure of the central part of the nebula.

The three-colour composite image was



taken last month during the telescope's first commissioning phase, in which bright objects were recorded. The second phase will look at faint objects.

Ehsan Masood

Senate approves bill to double research spending

[WASHINGTON] The Federal Research Investment Act, which would double research and development spending over a 12-year period, was passed unanimously by the US Senate on 9 October (see *Nature* **394**, 5; 1998).

The bill — first proposed by senators Bill Frist (Republican, Tennessee) and Jay Rockefeller (Democrat, West Virginia) — will not become law because the House of Representatives has not passed a companion measure. But its passage could improve the chances of the House considering such a measure next year, according to the scientific societies that have backed the proposal.

US Air Force faces legal challenge over chimps

[WASHINGTON] A group concerned about chimpanzee welfare is to sue the US Air Force over a decision taken in August to hand over its 111 research chimpanzees to the Coulston Foundation, whose treatment of other chimps has attracted the ire of animal rights activists.

The Center of Captive Chimpanzee Care,

whose board includes the primatologist Jane Goodall, says the New Mexico-based Coulston Foundation cannot meet the air force's own criteria for chimpanzee care, and will ask the Washington DC Court of Federal Claims to overturn the decision. Animal rights activists want the chimpanzees, which are left over from research associated with the space programme, to be put into a retirement sanctuary.

UN lab will investigate ocean pollution threat

[PARIS] The United Nations acquired its first oceanography laboratory last week with the opening in Monaco of the 3,000-square-metre Laboratory of the Marine Environment by the Vienna-based International Atomic Energy Agency (IAEA). The opening took place in the presence of Prince Rainier III and Mohamed El Baradei, the director general of IAEA.

The laboratory, which will have an annual budget of US\$5 million and more than 50 staff, will bring together the IAEA's various research activities in marine radioactivity and pollution with the aim of creating an international centre of excellence. The principality of Monaco has donated over \$10 million towards construction of the centre.

Senate halts appointment of FDA commissioner

[WASHINGTON] The US Food and Drug Administration (FDA) will have to operate without a commissioner at least until the new year, after a prominent senator said he would block the appointment of Jane Henney, President Bill Clinton's nominee. The move could derail the campaign against tobacco products energetically instigated by David Kessler, the previous commissioner.

An aide to Senator Don Nickles (Republican, Oklahoma), the majority whip in the Senate, told the *Washington Post* that Nickles and unnamed colleagues were concerned that Henney might attempt to regulate tobacco or allow the French abortion pill RU-486 to be prescribed in the United States. Henney is a former senior official at the FDA, and currently heads the University of New Mexico Medical Center.

NASA centre named after astronaut Glenn

[WASHINGTON] Congress has renamed the US space agency NASA's Lewis Research Center in Cleveland, Ohio, as the John H. Glenn Research Center, in honour of the Ohio-born astronaut-turned-

senator who later this month will return to space for the first time in 36 years. House and Senate appropriators directed the name change as part of the NASA funding bill expected to be signed into law this week.

Glenn, meanwhile, scolded news reporters last week for focusing too much on his age and celebrity rather than the "cutting edge" biomedical experiments that he and his crewmates will conduct in orbit. "I'm not back as a legislative passenger," Glenn insisted, "I'm back as a science passenger."

Brussels bid to block ban on modified crops

[PARIS] **The European Commission has warned the French government that its decision in July to introduce a two-year moratorium on the commercialization of some genetically modified crops, in particular maize, oil-seed rape and beetroot, is contrary to European Union (EU) legislation. This prevents member states from unilaterally blocking the commercialization of varieties of modified crops approved by the EU.**

The commission has also attacked the French decision to halt the cultivation of three varieties of modified maize

produced by Monsanto and AgrEvo. The decision was taken after the Conseil d'Etat, the highest administrative court in France, overturned an earlier authorization by the French government until it had studied the dossier in detail. A final decision is expected later this year.

Ig Nobel prizes spotlight Prozac-loving clams

[BOSTON] At the Eighth First Annual Ig Nobel Prize Ceremony held last week at Harvard University, Peter Fong of Gettysburg College in Pennsylvania captured the biology prize for "contributing to the happiness of clams by giving them Prozac". French researcher Jacques Benveniste received the chemistry award for the second time for his homeopathic discovery that the 'memory' of water can be transmitted over telephone lines (see *Nature* 395, 535; 1998).

Author Deepak Chopra won the physics prize for his unorthodox approach to quantum theory. Harvard physicist Roy Glauber challenged Chopra's views, arguing that "the successes of relativistic quantum mechanics in the fields of psychiatry and emotional well-being have been few".

But another Harvard physicist, Sheldon Glashow, was more charitable towards Chopra, asking: "Who else could have imagined quantum nutrition?"

A team of physicians from the Royal Gwent Hospital in Newport, Wales, won the medicine prize for a report published in *The Lancet* entitled "A man who pricked his finger and smelled putrid for 5 years".

British and Chinese academics get wired up

[LONDON] **An electronic network linking universities and colleges in the United Kingdom and China was established on 7 October under a scheme to improve communications between academic communities in the two countries. The network will link up the UK Joint Academic Network (JANET) with the Chinese Education and Research Network (CERNET), which at present connects up to 450 Chinese universities.**

The link is being funded by the UK and Chinese governments, as well as the communications company Cable and Wireless Communications. At present, Internet traffic between both countries has to pass through undersea lines and across the busy North American networks.

101 uses for a life sciences PhD

Sir — The underlying assumption of the report by the US National Research Council on PhDs in the life sciences is that leaving academic science is a disaster (*Nature* **395**, 103; 1998).

You say “the report dismisses the notion that PhD programmes should be broadened to prepare graduates for alternative careers that, it maintains, do not require a PhD”. This assumption that a life science PhD is only useful training for scientific research is narrow-minded and not dynamic. For many years it was said that manufacturing did not require a university degree. However, with many new university graduates seeking jobs in the 1980s, manufacturing was restructured to incorporate modern technology and now employs many university graduates. Thus, an apparent glut of overtrained people

increases productivity as their talents are transferred to new areas of employment.

We agree that providing better information to prospective students is an excellent idea. Applicants should know what jobs are available and what career paths exist both within and outside science. Then they should choose freely whether to begin their doctoral training. Setting an arbitrary quota on the number of PhDs allowed in the United States would be foolish. Economic history contains a wealth of examples of the failure of quota systems to perform efficiently. Young life scientists need improved information about career options, not limits on the number of PhDs.

Those who only want to clone themselves in economics departments are also lamenting that there are not enough jobs for PhDs at research universities. But

economics PhDs who go into consulting and investment banking do very well and they bring a willingness to entertain complex solutions and creative insights.

Life science PhDs have similar opportunities to use their skills in marketing, sales, management, and research in the biotechnology and pharmaceutical industries. Their technical training is also of use as biotech equity analysts, patent attorneys and venture capitalists. Who is to say we aren't improving the lives of the individuals and increasing productivity?

Andrew A. Bogan

Graduate Group in Biophysics,
University of California, San Francisco,
California 94143, USA

Elizabeth C. Bogan

Department of Economics, Princeton University,
Princeton, New Jersey 08544, USA

The Western front

Sir — As Native American university faculty members, we are concerned about a major error concerning Native American history in Adam Kuper's review of Mark Cocker's book *Rivers of Blood, Rivers of Gold* (*Nature* **393**, 533; 1998). Kuper's reference to the so-called Apache being “virtually bandit groups, established in the interstices between the British colonies and Mexico,” is both insulting and ignorant. Recent scholarship makes it clear that the Apaches were essentially agriculturalists, and their so-called raids were attempts to protect their homelands from invasion.

Although Kuper complains about Cocker's reversal of the old Victorian bias, it seems he has nothing better to offer than a return to even older stereotypes of bandit tribes and small nomadic bands. The historical reality was much more complex, and some emphasis should be made of the strong links between indigenous peoples and the places from which they come. Understanding this relationship defines the tragedies that Cocker describes.

Daniel Wildcat

Raymond Pierotti

Haskell Indian Nations University,
Lawrence, Kansas 66046, USA

Kuper replies — We are dealing here with very broad-brush characterizations, and the label ‘Apache’ covers a variety of pre-conquest lifestyles. However, it is generally agreed that farming was a minor aspect of pre-conquest Apache subsistence activities, and that raiding was very significant, particularly among the Western Apache.

According to *Encyclopedia Britannica*, the Chiricahua (Geronimo's division) “were perhaps the most nomadic and aggressive of the Apache west of the Rio Grande”. It was their nomadism that “made their predatory life-style possible”. It is true, nevertheless, that the equally aggressive settlers forced a number of Apache groups into heroic acts of self-defence which were wrongly represented as unjustified raids.

Adam Kuper

Department of Human Sciences,
Brunel University,
Uxbridge,
Middlesex UB8 3PH, UK

Coelacanth populations may go with the flow

Sir — The coelacanth population recently discovered in the Sulawesi Sea, north of Sulawesi, Indonesia^{1,2}, may very well be related to those found earlier near the Comoran Islands in the western Indian Ocean. The ocean circulation map shown² is in error: there is indeed an oceanographic connection between Sulawesi and the Comoran Islands region.

Oceanographic studies^{3–5} show that there is a direct flow of surface and thermocline (upper 400 m) water from the Mindanao Current of the North Pacific into the Sulawesi Sea. After traversing Makassar Strait this flow, called the Indonesian throughflow, enters the Flores and Banda Seas before spreading into the Timor Sea and Indian Ocean. Once in the Indian Ocean the Indonesian water spreads

westward as a low salinity feature embedded within the South Equatorial Current along 12° S, eventually reaching the vicinity of the Comoran Islands. I suggest that coelacanth populations may exist along the route of this flow.

Arnold L. Gordon

Lamont-Doherty Earth Observatory,
Palisades, New York 10964-8000, USA

1. Erdmann, M. V., Caldwell, R. L. & Moosa, M. K. *Nature* **395**, 335 (1998).
2. Forey, P. *Nature* **395**, 319–320 (1998).
3. Gordon, A. L. *J. Geophys. Res.* **C 91**, 5037–5046 (1986).
4. Gordon, A. L. & Fine, R. A. *Nature* **379**, 146–149 (1996).
5. Gordon, A. L. *et al.* *Geophys. Res. Lett.* **24**, 2573–2576 (1997).

Launch editor

Sir — Robert Muir-Wood's highly readable review of *Annals of the Former World* by John McPhee (*Nature* **394**, 845–846; 1998) includes an error about one of McPhee's geological guides: Eldridge Moores was indeed a powerful force in shaping the Geological Society of America's journal *Geology*, but he did not found it.

Geology was launched by edict of the society's council in 1973 and was shepherded by Bennie Troxel, then the society's science editor. Henry Spall (of the US Geological Survey) was appointed *Geology* editor and guided the journal through its early years. Moores, as editor from 1982 to 1988, applied his considerable energy to expanding *Geology*, both in subject matter and size. Moores is, however, erudite and charming, as Muir-Wood says.

Faith Rogers

Geological Society of America, 3300 Penrose Place,
PO Box 9140, Boulder, Colorado 80301, USA

How big do stellar explosions get?

Eddie Baron

We thought we knew how powerful supernova explosions could be. We also thought that supernovae and γ -ray bursts were unrelated. One extraordinary supernova is making us re-examine these ideas.

Supernovae are the explosive death throes of massive stars, and they produce the elements that the Earth and you and I are made of. In a 'core-collapse' supernova — one of two main types — the star's core becomes so large that it collapses. The mass of electrons and nuclei becomes compressed into a neutron star, a tiny ball of

neutrons about ten kilometres across; that releases a vast amount of gravitational energy into the surrounding material, blowing it outwards. For a few weeks a single star becomes as bright as a galaxy of 100 billion ordinary stars. Supernovae release their energy in many forms: neutrinos, the kinetic energy of the debris, optical light and radio waves.

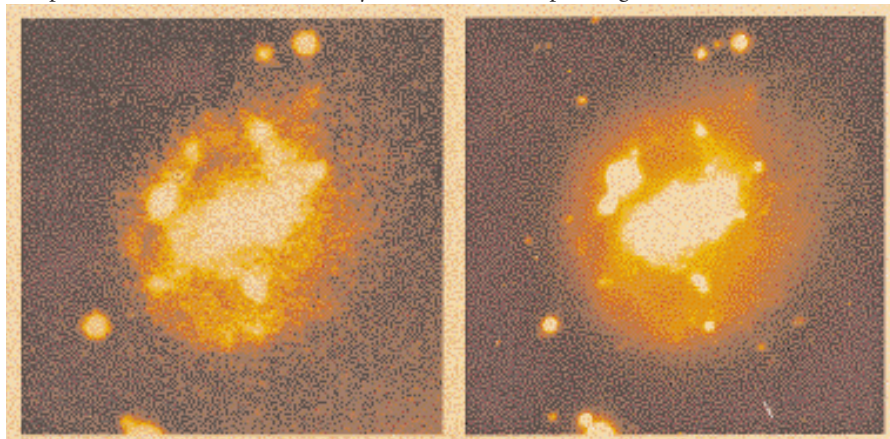


Figure 1 **Before and during:** the galaxy ESO184-G82 in its normal state (left) and with supernova 1998bw going off in one of its spiral arms (right).

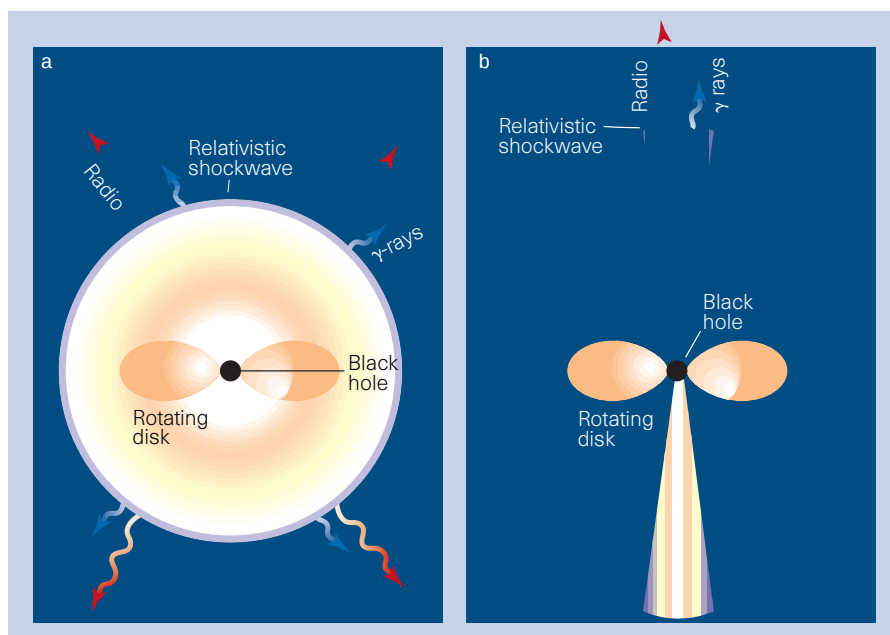


Figure 2 **Explosive shapes.** a, If SN1998bw was spherical, it must have been about 30 times as powerful as an ordinary supernova. b, Less energy is needed for an explosion in two narrow cones. But in both cases, the strangeness of this supernova and the coincident γ -ray burst seem to imply that the star's core collapsed directly into a black hole.

In contrast, γ -ray bursts are enigmatic explosions that produce nearly pure bursts of very energetic light — γ -rays. This radiation may be produced by an explosion shock wave travelling at nearly the speed of light, but we don't know what causes the explosion in the first place. We do know that they are powerful: from their uniform distribution on the sky, most bursts must occur at cosmological distances (billions of light years away) and for at least one burst such a large distance has been measured directly¹. So to appear as bright as they do, bursts must be intrinsically very luminous, putting out even more energy than a supernova in just a few seconds.

A supernova occurs about once a second in the observable Universe, a γ -ray burst about once a day. That a supernova and a γ -ray burst could be produced by the same event would be an extraordinary revelation and would shed light on just what objects are producing γ -ray bursts. Three papers in this issue²⁻⁴, on pages 663, 670 and 672, argue that some events can indeed produce both a supernova and a burst.

Supernova 1998bw occurred at about the same time and in the same region of the sky (Fig. 1) as the γ -ray burst (GRB) 980425. Galama *et al.*³ discuss the probability that the supernova and the γ -ray burst are the same object. The two detections were within a few days of each other, and within a few degrees, suggesting a connection. Unfortunately, the arrival direction of γ -rays cannot be determined by current detectors to better than a few degrees, but in some cases X-ray afterglows to γ -ray bursts have been seen, and pinned down to within a few arc minutes. Several X-ray sources were detected² near SN1998bw, but their positions do not agree with that of the supernova. In any case, although it is possible that there was a chance coincidence between the supernova and the burst, why should nature capriciously choose what was a particularly odd supernova — exceedingly energetic, exceptionally luminous and with peculiar radio emission?

At radio wavelengths it is the brightest supernova ever observed, according to Kulkarni *et al.*². In all other radio supernovae seen before, the source slowly brightens over time, taking weeks to years to reach peak brightness, with the brighter ones peaking later. But the radio emission from SN1998bw came very early, within ten days of the explosion. Kulkarni *et al.* argue that the radio emission must be produced by electrons moving at nearly the speed of light, and that it is therefore physically connected with the γ -ray burst.

If GRB980425 was indeed coincident with the supernova it must have been very weak. That is because SN1998bw occurred in a spiral galaxy only about 140 million light years away. GRB980425 didn't appear particularly bright as γ -ray bursts go, so if it was so

close it must have been intrinsically dim — perhaps a member of a new class of dim, supernova-related γ -ray bursts^{5–7}.

But what sort of event could appear to us as both an odd supernova and an odd γ -ray burst? Iwamoto *et al.*⁴ use a model in which they assume the core of a very massive star collapses into a black hole, rather than a neutron star. Thirty times more kinetic energy is generated than in an ordinary supernova, and the predicted light curve (brightness versus time) agrees well with observations. Such events have been dubbed ‘collapsars’⁸ and ‘hypernovae’⁹.

One problem with this idea as an explanation for SN1998bw is that it is hard to explain how the γ -ray burst was produced. Even a small proportion of protons and neutrons in the outer shock wave will slow it down to the point at which few γ -rays are generated, although there could be a very fast outer shell to the explosion.

Perhaps one can produce a faster shock wave⁴ if the explosion has a special orientation to our line of sight^{5,10}, expanding violently along a narrow cone pointed towards us (presumably at one pole of the progenitor star). Even more importantly, an asymmetric explosion would not need to have such a huge total energy. If explosions such as this are highly asymmetric, when we happen to be looking right down the symmetry axis we might see both a γ -ray burst and a supernova; but in most cases we wouldn’t happen to be in that position and so we would miss the narrow cone of γ -rays, and the supernova would appear less remarkable (Fig. 2).

Unfortunately, Kulkarni *et al.* do not see a preferred direction to the radio emissions, and conclude that we are looking at a nearly spherical object. These results seem to be confirmed by observations at optical wavelengths^{11,12}, but until detailed numerical simulations are performed that conclusion is not definitive.

Even without the γ -ray burst, SN1998bw is a strange supernova. By modelling its light curve we can estimate the supernova’s kinetic energy, the total mass ejected and the amount of ^{56}Ni mixed in (this radioactive isotope warms the supernova as it decays, making the remnant fade more slowly). In the standard picture of core-collapse supernovae, the explosion is nearly spherical: around $10M_{\odot}$ (one $M_{\odot}$ is the mass of the Sun) of material is ejected with a kinetic energy of about 10^{51} erg, and about $0.1M_{\odot}$ of radioactive ^{56}Ni . But this doesn’t fit SN1998bw. Its explosion can be modelled as either the spherically symmetric explosion of a very massive star, with about $0.5M_{\odot}$ of ^{56}Ni in a total ejected mass of $\sim 10M_{\odot}$ with $\sim 30 \times 10^{51}$ erg of kinetic energy^{3,6}; or as an asymmetric explosion ejecting $\sim 2M_{\odot}$ of material with $\sim 2 \times 10^{51}$ erg of kinetic energy, and $\sim 0.2M_{\odot}$ of ^{56}Ni (ref. 13). Both of these alternatives are difficult to reconcile with the standard picture — a

picture verified by observations of supernova 1987A, which occurred only about 180,000 light years away.

The spectrum of SN1998bw is similar to that modelled by my colleagues and me for another peculiar supernova, SN1997ef, where a high kinetic energy is also inferred. However, SN1997ef appears not to be associated with a γ -ray burst¹⁴.

All of this may lead to a revolution in our thinking about how core-collapse supernovae are produced. Do ordinary ones like SN1987A happen only in stars with a mass of about $\sim 20M_{\odot}$, leaving more massive stars of $\sim 40M_{\odot}$ to produce either very energetic or very asymmetric explosions? Or is there a kinetic-energy limit for core-collapse supernovae of 10^{51} erg, but sometimes the ejecta is spherically distributed and at other times it is not. And when does a γ -ray burst accompany the collapse?

We won’t know until we can determine whether SN1998bw was spherical or not. To do that, observations are not enough — they must be interpreted in the light of theoretical

models. Unfortunately, our most detailed models at present are limited to the simplest case of spherical symmetry by the speed of our fastest computers. With advances in both computer power and numerical techniques in the next millennium, detailed three-dimensional calculations may finally tell us how big an explosion a star can make. □

Eddie Baron is in the Department of Physics and Astronomy, University of Oklahoma, Norman, Oklahoma 73019, USA.

e-mail: baron@mail.nhn.ou.edu

1. Metzger, M. R. *et al.* *Nature* **387**, 878–880 (1997).
2. Kulkarni, S. R. *et al.* *Nature* **395**, 663–669 (1998).
3. Galama, T. J. *et al.* *Nature* **395**, 670–672 (1998).
4. Iwamoto, K. *et al.* *Nature* **395**, 672–674 (1998).
5. Wang, L. & Wheeler, J. C. *Astrophys. J.* **504**, L87–L90 (1998).
6. Woosley, S. E., Eastman, R. & Schmidt, B. *Astrophys. J.* (submitted).
7. Bloom, J. *et al.* *Astrophys. J.* (submitted).
8. Woosley, S. E. *Astrophys. J.* **405**, 273–277 (1993).
9. Paczyński, B. *Astrophys. J.* **494**, L45–L48 (1998).
10. Cen, R. *Astrophys. J.* (submitted).
11. Patat, F. & Piemonte, A. *IAU Circ. No.* 6918 (1998).
12. Kay, L., Halpern, J., Leighly, K. & Heathcote, S. *IAU Circ. No.* 6969 (1998).
13. Höflich, P., Wheeler, J. C. & Wang, L. *Astrophys. J.* (submitted).
14. Norris, J., Bonnell, J. & Watanabe, K. *Astrophys. J.* (submitted).

Human evolution

Towards a genetic history of China

Alberto Piazza

Scholars of human evolution are well aware that physical traits are mostly adaptive, reflecting the geography or environment — rather than the history — of those who carry them. Therefore, they will appreciate a paper by Chu and thirteen other Chinese co-authors¹ in *Proceedings of the National Academy of Sciences*, who have done a genetic analysis of China at a finer resolution than any attempted so far.

In his book *Guns, Germs, and Steel*², Jared Diamond wonders how China — politically

unified since 221 BC and linguistically almost monolithic, with a single writing system from the beginning of its recorded history — happened to become Chinese. One reason why we should not have expected such astonishing homogeneity (which covers one-fifth of our present species and one-fifteenth of our inhabited planet) is genetic. Northern and southern Chinese people, whose boundary is conventionally located between the Huang Ho (Yellow River) and the Yangtze, were previously found to be

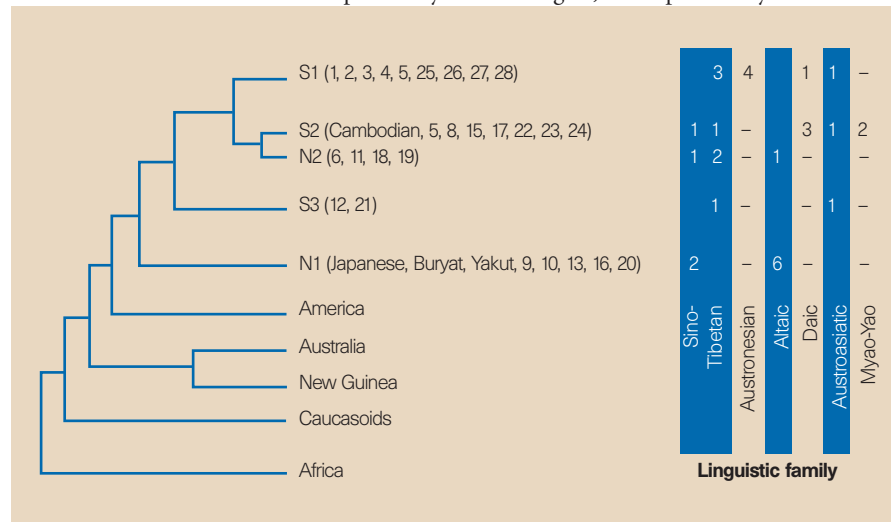


Figure 1 Main features of the tree models constructed by Chu *et al.*¹. The authors sampled 28 groups from all over China, and compared microsatellite DNA markers from them with samples from non-Chinese control populations. The southern (S) group of Chinese samples were grouped into three clusters, and the northern (N) group into two.



Figure 2 Map of China showing locations of the sampled populations, and geographical display of a phylogenetic tree based on Han surnames (from ref. 9). Sampled populations are numbered 1–28, but sample 7 is not shown as it is Han from the Bay area of California.

genetically different — at least from the relatively few samples tested by blood-group and protein-marker polymorphisms. But this heterogeneity, supported by anthropological, environmental, archaeological, pre-historic and historical evidence, was far from a robust result derived from well-planned genetic surveys at a micro-geographical level. Physical appearance, as anywhere and at any time, was an important factor in recognizing these differences. The northern Chinese tend to be taller and paler, with smaller eyes that seem more 'slanted'. Moreover, old skeletons differ, the southern ones being more similar to contemporary South-east Asian types, especially those with a darker skin (such as the so-called Negritos)³.

Chu *et al.*¹ used 15 to 30 microsatellites to test genetic variability in 28 samples from the different provinces of China. Microsatellites are repeats of short DNA segments, widely distributed in our genome, that are commonly used in evolutionary analyses because they are numerous, inherited in Mendelian fashion, highly variable and easy to handle. The Chinese government officially recognizes 56 different ethnic groups. The Han (from the name of the great Han dynasties, 206 BC to AD 220) is the most important in terms of numbers (1.1 billion people) and historical tradition. The other 100 million Chinese people are found mostly in southern China, nearly half of them in one province, Yunnan.

Chu and colleagues sampled four Han groups and 24 minority groups. They also tested four East Asian, two American Native, one Australian, one New Guinean, four Caucasian and three African samples with the

same set of microsatellites, as non-Chinese control populations. They summarize the genetic differences in two phylogenetic trees, the main features of which are shown in Fig.

Behavioural ecology

Pick out a penguin

It's party time. The hi-fi is booming, and people are chattering loudly. Yet you can readily hear a couple talking at the other side of the room. This is the 'cocktail party effect', a phenomenon familiar to attentional psychologists. And penguins.

Breeding colonies of king penguins (*Aptenodytes patagonicus*) on subantarctic islands can number up to 300,000 birds. Penguin chicks, of course, need feeding, and must find the parent that has been foraging at sea on the latter's return. It is known that the initial detection is by vocal cues rather than sight or smell. But how good are chicks at identifying the parent's call when it is masked by the background din of the colony and screened by intervening bodies?

T. Aubin and P. Jouventin have tackled the question (*Proc. R. Soc. Lond. B* 265, 1665–1673; 1998). They measured the amplitude and other acoustic properties of parental calls and the ambient noise of a king-penguin colony; they then assessed the propagation of the call in the centre of the feeding area, as compared to an open area, to quantify the screening effect of bodies. In playback experiments, a chick

1. This tree is robust (as measured by the 'bootstrap' reproducibility test⁴) only where it represents populations that are very distant from one another both geographically and genetically (that is, populations outside East Asia). There are much smaller genetic differences between East Asian populations, and historical records document many incidents of possible gene flow between them. So a tree model of evolution may not be the best one to describe the genetic history of China — as has been repeatedly pointed out in the genetic analysis of European populations⁵.

The main structure of the tree agrees closely with previous results using classical, non-DNA genetic markers⁶. Its root separates African from non-African populations, and all East Asian populations cluster together, their nearest genetic neighbours being American Natives, followed by Australian aborigines and New Guineans. These results agree with the subsequent settlement times of Australia (about 60,000 to 50,000 years ago) and the Americas (from 30,000 to 15,000 years ago). The southern group of Chinese samples is distributed in three genetically related clusters, called S1, S2 and S3 (Fig. 1). The clusters differ in the number of minority groups that inhabit the Yunnan region, and in the language distribution; only one Han Chinese-speaking group is included (in S2) from the province of Henan.



reacted only to calls made by its own parent, turning and then running towards the source. Further experiments involved 'jamming' the calls of one parent by mixing in the calls of other adults, mimicking the true situation in the feeding zone.

The authors' calculations suggest that the maximum distance at which the chick should be able to detect its parent against the background cacophony should not exceed 8–9 metres. But even when the parent's calls were jammed, chicks picked them out at nearly twice that distance. This remarkable feat of auditory discrimination is yet to be explained.

Rory Howlett

The northern samples are distributed in two clusters, N1 and N2. The N1 cluster includes six Altaic-speaking samples, one northern Han sample and one Han sample from the Yunnan province. The N2 cluster, which includes four minority groups (only one of them, from the Ninxia province, is Chinese speaking), has historically well-established northern origins. Unexpectedly, it associates in the tree with the southern S2 cluster. Chu *et al.*¹ think it possible that the populations sampled in N2 mixed more with southern populations. But only more microsatellite markers will tell us whether this N2–S2 association is spurious, or whether the tree is not the appropriate model to represent ethnic groups that were heavily exposed to admixture in the past.

What messages can we draw from Chu and colleagues' analysis? The problem of the origin of northern and southern East Asians can be summarized by three different models. The first postulates that populations from the northeast migrated to the south, and mixed with the 'Australoids' (Australian aborigines and Papuans) who had settled in Southeast Asia. This generated (hypothetically, 30,000 to 50,000 years ago) the 'hybrid' group of southern East Asians. The second model assumes that the northern populations evolved from the southern East Asian ones. And a third model assumes that northern and southern East Asian populations have evolved independently since the late Pleistocene (more than 10,000 years ago)⁷.

These models could be compatible, depending on the time window considered. The first is consistent with the linguistic map of China. We can infer from it that north China was formerly occupied by speakers of Chinese and other Sino-Tibetan languages, whereas parts of south China were variously occupied by speakers of Austro-Asiatic, Austronesian, Miao-Yao and Daic languages. Latterly, Sino-Tibetan (during the Chinese Zhou dynasty, about 3,000 years ago⁸) replaced most of the other linguistic families over south China, by a similar process to the diffusion of Indo-European speakers over Europe⁵.

The third model is consistent with the archaeological record, which places at least three independent centres where agriculture and associated Neolithic cultural areas originated in China. These are in the north (Yang-Shao culture, middle Huang Ho river, beginning about 8,000 years ago), where millet was the main crop and pigs and dogs were domesticated; on the southern coasts, including Taiwan (Ta-Pen-Keng culture, also beginning around 8,000 years ago), where the main product was rice; and on the lower Yangtze river (Qing-Lien-Kang culture, beginning 1,000 years later), where rice was also cultivated and pigs and dogs were domesticated. Interestingly enough, the geographical distribution of these

Neolithic cultures is reflected in the geographical distribution of present Han surnames⁹ (Fig. 2). Surnames are very useful and efficient 'genetic' markers, transmitted only by the male line. In China they go as far back as 4,000 years, to the Xia dynasty (about 2100–1600 BC), and most of them predate the major northern invasions.

Finally, the second model is consistent with the genetic data presented by Chu *et al.*¹. Other genetic⁶, dental¹⁰ and cranial¹¹ data also support the implied hypothesis, although they are not incompatible with the third model. The difference is probably the time span to which they refer. Chu and colleagues conclude that "modern humans [that] originated in Africa constitute the majority of the current gene pool in East Asia", intending to falsify the hypothesis of a continuity of evolution from *Homo erectus* to anatomically modern humans in East Asia¹². The rationale is that their tree indicates it more likely that ancestors of the Altaic-speaking people sampled in East Asia entered Asia from Southeast rather than from Middle Asia. I am persuaded that this suggestion is correct, but the lack of samples from Middle Asia still makes it little more

than a working hypothesis. In any case, the collaboration set up by Chu *et al.* is a remarkable step towards a deeper biological and cultural understanding of this most important part of the world. □

Alberto Piazza is in the Dipartimento di Genetica, Biologia e Biochimica, Università di Torino, via Santena 19, 10126 Torino, Italy.
e-mail: a.piazza@cios.to.cnr.it

1. Chu, J. Y. *et al.* *Proc. Natl Acad. Sci. USA* **95**, 11763–11768 (1998).
2. Diamond, J. *Guns, Germs, and Steel: The Fates of Human Societies*, Ch. 2 (Norton, New York, 1997).
3. Chang, K. C. *The Archaeology of Ancient China* (Yale Univ. Press, New Haven, 1977).
4. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman & Hall, New York, 1993).
5. Piazza, A. *Science* **260**, 1767–1769 (1993).
6. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton Univ. Press, 1994).
7. Omoto, K. in *The Origin and Past of Modern Humans as Viewed from DNA* (eds Brenner, S. & Hanihara, K.) 92–109 (World Scientific, Singapore, 1995).
8. Wang, W. S.-Y. *Explorations in Language* 375–396 (Pyramid, Taipei, 1991).
9. Du, R., Yuan, Y., Hwang, J., Mountain, J. & Cavalli-Sforza, L. L. *J. Chin. Ling.* **5**, 1–93 (1992).
10. Turner, C. G. II *Am. J. Phys. Anthropol.* **82**, 295–317 (1990).
11. Hanihara, T. *Anthropol. Sci.* **101**, 25–46 (1993).
12. Wu, X.-Z. in *The Origin and Past of Modern Humans as Viewed from DNA* (eds Brenner, S. & Hanihara, K.) 267–282 (World Scientific, Singapore, 1995).

Cosmology

New troubles for inflation?

Marc Kamionkowski and Andrew H. Jaffe

In the very early Universe, gravity may briefly have become a repulsive, rather than an attractive, force. The ensuing period of ultra-rapid expansion — called 'inflation'^{1–3} — could account for some of the fundamental features of the Universe, such

as the remarkable smoothness of the 2.7 K cosmic microwave background (CMB) radiation and the origin of large structures such as galaxies and galaxy clusters. But an analysis by Ferreira, Magueijo and Gorski, published in *The Astrophysical Journal Letters*⁴,

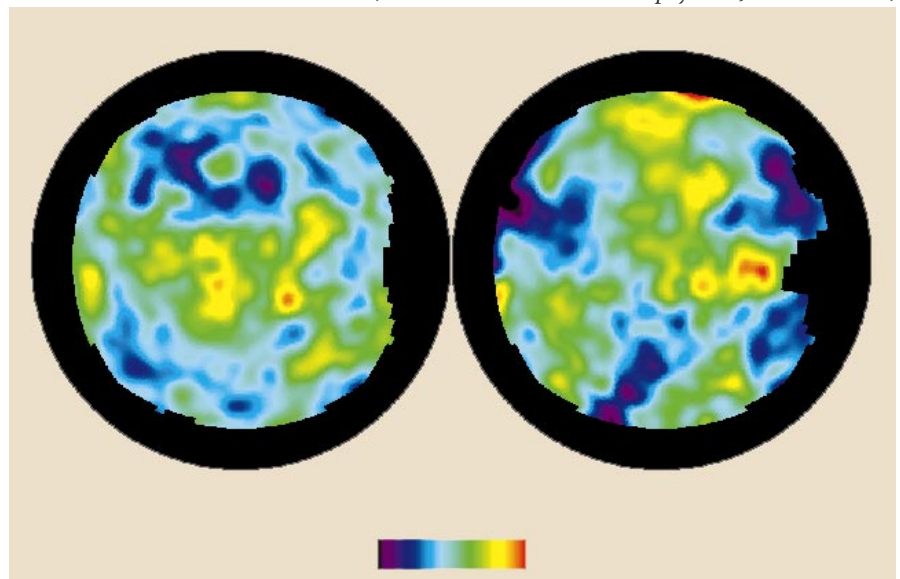


Figure 1 A map of the Universe when it was just 300,000 years old. Slight fluctuations in density show up as temperature variations in the present-day microwave background, here measured by the Differential Microwave Radiometer instrument on board the COBE satellite. There is a hint of peculiar behaviour in these data — a correlation between different points that is not predicted by the theory of inflation.

NATURE

100 YEARS AGO

A complimentary dinner was given to Prof. Virchow at the Hôtel Métropole on Wednesday in last week. ... Lord Lister, in proposing the toast of the evening, dwelt upon the versatility of the genius of the distinguished guest, his eminence as a pathologist being equalled by his reputation as an anthropologist and antiquarian. He referred particularly to Virchow's "Cellularpathologie," which work, he remarked, "swept away the false and barren theory of a structureless blastema, and established the true and fertile doctrine that every morbid structure consists of cells which have been derived from pre-existing cells as a progeny. Cellular pathology is now universally recognised as a truth. Even those morbid structures which deviate most from the normal structure are known to be derived as a progeny from normal tissue – from normal cells, driven to abnormal development by injurious agencies."

From *Nature* 13 October 1898.

50 YEARS AGO

"Einstein: His Life and Times." By Philipp Frank. – Those who, like the present reviewer, are personally unacquainted with Einstein, will read this book with a shock of surprise. While Dr. Frank's sympathies are all with Einstein, the portrait presented to us is not altogether a pleasant one. We see a man developing early into the traditional type of nineteenth-century 'professor'. He regards himself as free to develop any eccentricity of behaviour, whether those about him like it or not, and to talk shop in season or out of season, a characteristic illustrated by the description of a courtesy visit to a non-mathematical colleague in Berlin, in which, after subjecting his hosts to a forty-minutes discourse on relativity, Einstein left abruptly. This lack of appreciation of the fact that ideas and interests which did not happen to interest him might still be as valuable as those that did may well explain Einstein's difficulties in the Berlin Academy, or his failings as a teacher. Always, apparently, ready to lecture on his researches of the moment or to deliver popular discourses, Einstein was not prepared to teach his students systematically what they had need to learn. – G. C. McVittie

From *Nature* 16 October 1948.

may put a dent in inflation's armour. The inflationary models favoured by theorists are lauded for the simplicity of their predictions, but if these new results withstand further scrutiny then inflation either cannot be so simple, or must be discarded.

Among inflation's predictions is that the distribution of density perturbations in the early Universe should be Gaussian (or 'normal'): a histogram of the density at different points in the Universe should produce the familiar bell-shaped curve. Gaussianity also implies specific forms for the correlation between densities at several different points. The new work provides preliminary evidence that contradicts these predictions.

The Differential Microwave Radiometer experiment on the Cosmic Background Explorer (COBE) satellite has mapped the temperature of the CMB with an angular resolution of seven degrees⁵ (Fig. 1). The photons come to us from roughly 15 billion light years away; they were produced about 300,000 years after the Big Bang, when the Universe became transparent and the photons last scattered from electrons in the primordial cosmic plasma. Photons from overdense regions of the Universe lost more energy as they climbed out of deeper gravitational potential wells and are thus redder than photons reaching us from underdense regions.

The distribution of temperature fluctuations thus reflects the primordial distribu-

tion of mass, and the observed temperature fluctuation level — roughly one part in a hundred thousand — reflects a similar level of density fluctuations. The distribution appears at first to be roughly consistent with a Gaussian. But there are only a small number of independent 7×7 -degree regions on the sky, so there is an error in each point in a histogram constructed from the data, simply from sample variance. Detector noise provides another source of uncertainty. There must, therefore, be some deviations from Gaussianity. Are they statistically significant, or just sampling and noise fluctuations?

Several searches have yielded null results. But there are an infinitude of flavours of non-Gaussianity: a distribution might be wider or narrower, have longer tails than a Gaussian, be skewed and so on. Still more flavours are allowed when correlations between regions are considered — and any deviation in the histogram from a perfect bell curve poses problems for inflation.

Ferreira *et al.*⁴ have developed a technique to test for a specific form of non-Gaussianity. They search for this particular correlation between three points on the sky. When applied to the data from the Differential Microwave Radiometer, the search turns up a positive result at the 98% confidence level. As it probes a different form of non-Gaussianity, the result is perfectly consistent with prior null results.

Photonics

Chemistry of light

This picture is of a 'photonic molecule', the latest development in our increasingly sophisticated control of photons. Bayer and colleagues (*Phys. Rev. Lett.* 81, 2582–2585; 1998) have built a range of these structures to test how pure modes of trapped light behave in them.

Until a few years ago, our manipulation of light was crude: creating photons of visible light by heating matter until its electrons vibrate violently enough to radiate; filtering them; bouncing them off mirrors. The light is generally incoherent, and has a wide range of colours. Lasers are a much more refined tool, producing a narrow beam of nearly monochromatic, coherent (in-phase) light.

Even cleverer are photonic-bandgap materials, where scattering off periodic arrays of a dielectric prevents certain wavelengths of light from propagating. Such materials can be used to bend light around corners, and trap it in cavities.

Light can also be trapped in 'photonic atoms'. A semiconductor layer is sandwiched between two mirrors, and an island is etched out of the overall structure. The refractive-index barrier between the



semiconductor and the surrounding air allows photons to leak only slowly out of the sides; this confinement means that only resonant photon modes are excited. This technology may be used to make efficient semiconductor lasers with a single excited mode.

To take the idea a stage further, Bayer *et al.* linked pairs of gallium-arsenide photonic atoms with narrow bridges. As the bridge length was reduced, new photonic states emerged that are similar to the bonding and antibonding orbitals of the hydrogen molecule.

The aim now is to build more complicated photonic molecules to study this photon-matter interaction, and perhaps to learn something, by analogy, about electronic states in real molecules.

Stephen Battersby

Given that these results would rule out inflation as we know it, they require close scrutiny. Cosmologists will first wonder about systematic effects that must be taken into account in reconstructing the CMB temperature from the raw data. The authors have already shown that the most obvious of these — the geometric effects from removing regions of the sky contaminated by galactic emission — are not important. They also argue that their result cannot be a statistical fluctuation from having a finite number of independent lines of sight. This is particularly interesting, as it shows that this and future searches for non-Gaussianity will not be limited by such 'cosmic variance'.

Might deviations from Gaussianity in detector noise be responsible? The authors say no: they measured the instrument noise and took it into account. But it may be that their noise model, based on prior measurements, leaves out the particular deviation from Gaussianity they infer in the cosmological data.

Another possibility is that foreground contaminants, such as radio emission from our Galaxy, are responsible. That would require a larger degree of contamination from unidentified sources than previously estimated, having important implications for foreground-subtraction algorithms used for COBE and those being developed for future CMB experiments.

If the reported non-Gaussianity is real, it will pose serious problems for structure-formation theories, as inflation seems to be the only game in town. That is not for lack of imagination. Less than a decade ago, theorists had a plethora of plausible models such as topological defects (cosmic strings or textures), explosions and phase transitions; and more bizarre possibilities such as a 'loitering' Universe and some sort of density 'seeds'. With COBE, virtually all of these alternatives became unlikely, at least in their simplest varieties. But deviations from Gaussianity might resurrect some of these models, modified to fit available data. Or perhaps non-Gaussianity is a consequence of some more complicated theory of inflation⁶, making these results a window onto the high-energy physics of the early Universe.

Again, if confirmed, these results will raise new questions. What underlying pattern of temperatures could produce this signature? What physical process might be responsible? The analysis has detected a 'spectral line' of non-Gaussianity, fluctuations on an angular scale not much larger than the instrument's beam size. So does it continue to smaller scales? Is the signal limited to a small area of the sky? Further measurements could provide answers. For example, topological defects would produce coherent structures in higher-resolution maps, but some funny model of inflation would not.

Wrong or right, these results are provocative and worth pursuing. Fortunately, a satellite being developed by NASA called the Microwave Anisotropy Probe (MAP) and one by ESA called the Planck Surveyor, should show whether the conclusions of Ferreira *et al.* are correct. And Gaussianity (or lack thereof) is just one of many predictions of structure-formation theories. Each model must reproduce the observed large-scale structure of the Universe, and makes a specific prediction for the angular power spectrum of the CMB temperature and its polarization, and these will be measured by MAP and Planck with unprecedented precision. □

Marc Kamionkowski is in the Department of Physics, Columbia University, 538 West 120th Street, New York, New York 10027, USA.

e-mail: kamion@phys.columbia.edu

Andrew H. Jaffe is at the Center for Particle Astrophysics, University of California, 301 Le Conte, Berkeley, California 94720, USA.

e-mail: jaffe@cfpa.berkeley.edu

1. Guth, A. H. *Phys. Rev. D* **28**, 347–356 (1981).
2. Linde, A. D. *Phys. Lett. B* **108**, 389–393 (1982).
3. Albrecht, A. & Steinhardt, P. J. *Phys. Rev. Lett.* **48**, 1220–1223 (1982).
4. Ferreira, P. G., Magueijo, J. & Gorski, K. M. *Astrophys. J. Lett.* **503**, L1–L4 (1998).
5. Smoot, G. F. *et al.* *Astrophys. J.* **371**, L1–L4 (1991).
6. Salopek, D. S. *Phys. Rev. D* **45**, 1139–1157 (1992).

Developmental biology

Cripto-analysis of embryonic codes

Rosa Beddington

One of the earliest tasks for any embryo is to decide where to form its head as opposed to its tail. This establishes the antero-posterior axis of the organism. In *Drosophila* embryos, this axis is dictated by the molecular asymmetries that are generated during development of the egg. But in vertebrate embryos, such as the frog, the axis is thought to result from the influence of the organizer — a specialized region that forms on one side of the embryo after fertilization. Products of the organizer cause the ectoderm on the surface of the frog embryo to become neural, and give this neural tissue an explicit antero-posterior pattern¹. However, studies on the mouse

embryo, including those described by Ding *et al.*² on page 702 of this issue, indicate that the time-honoured organizer may not determine where the head forms in mammals.

Mammalian embryos develop inside the mother, and their first priority is to make the extra-embryonic tissues that are necessary for attaching to the uterus and obtaining nutrients. By the time the mouse embryo is patterned, including definition of its antero-posterior axis, the conceptus is almost a week old. By this stage it has implanted in the uterus and contains several layers of extra-embryonic tissue. These envelop a single sheet of epithelium, the

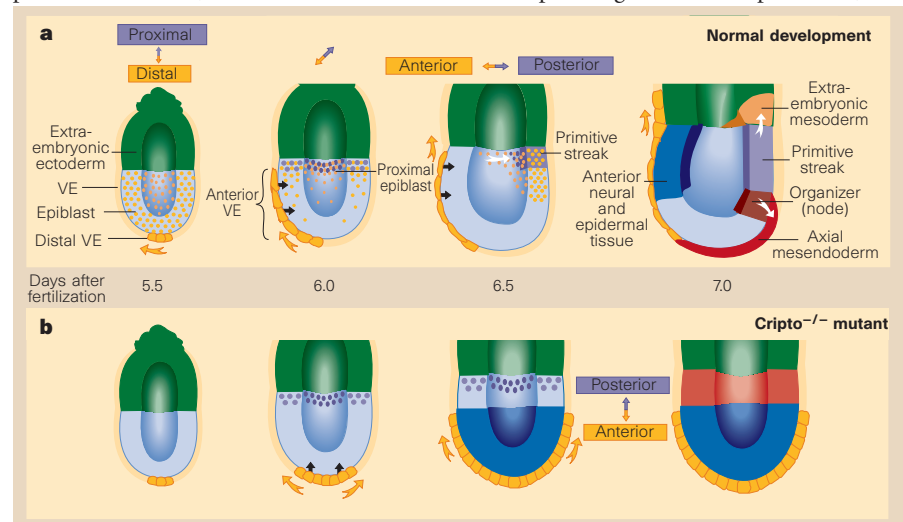


Figure 1 Antero-posterior patterning in normal and *Cripto*-mutant mouse embryos. a, Normal mouse development, showing the conversion of proximo-distal asymmetry into the definitive antero-posterior axis. (The most external extra-embryonic tissue layer is omitted.) Expression patterns are shown for *Hex* (orange) in visceral endoderm (VE), primitive-streak markers (purple), *Cripto* (yellow) and markers of anterior tissue (dark blue). Direction of VE movement is indicated by orange arrows, and movements of epiblast and its derivatives by white arrows. Black arrows indicate induction of anterior identity. b, As Ding *et al.*² now show, *Cripto* null mutants produce anterior and posterior tissue, but they cannot orientate the antero-posterior axis.

epiblast, which alone is destined to form the entire mouse (Fig. 1a). The future posterior end of the embryo is defined by formation of the primitive streak, an avenue through which epiblast cells can move to form a new tissue layer, the mesoderm. The anterior end of the streak corresponds to the organizer which, in the mouse, has many of the inductive and gene-expression properties of its frog counterpart. It gives rise to the axial mesendoderm, which populates the midline of the embryo. The middle of the streak produces more lateral embryonic mesoderm, whereas extra-embryonic mesoderm emerges from its posterior end. Traditional thinking is that axial mesendoderm from the organizer should induce anterior character in the embryo.

Using targeted mutagenesis, Ding *et al.*² studied the function of *Cripto*. This gene belongs to a newly discovered family that encodes secreted proteins with cysteine-rich and epidermal-growth-factor-like motifs. *Cripto* is initially expressed exclusively and uniformly in the epiblast. Then, gradually, its transcripts become restricted to the proximal rim of the epiblast, which is fated to form the primitive streak (Fig. 1a). The authors found that embryos lacking *Cripto* have neither a streak nor an embryonic mesoderm, although extra-embryonic mesoderm is present. Markers of the streak are transiently expressed throughout proximal epiblast, but they never become localized — as they should if a streak formed. Likewise, the organizer and axial mesendoderm are absent. Yet in spite of this, the epiblast differentiates into neural tissue that shows only anterior (forebrain and midbrain) character (Fig. 1b). How can this anterior identity arise without an organizer?

It transpires that antero-posterior asymmetry in the mouse embryo not only precedes formation of the streak — and hence the organizer — but that it first appears extra-embryonically in the visceral endoderm. Immediately after the embryo has implanted, expression of a homeobox gene, *Hex*, delineates a small cluster of visceral endoderm cells at the distal end of the embryo (Fig. 1a). Expression of *Hex* then shifts to one side of the embryo because, remarkably, these distal visceral endoderm cells give rise only to anterior progeny³. Thus, almost a day before the primitive streak is formed, there is asymmetrical cell deployment in the visceral endoderm, converting a proximo-distal difference in gene expression to an antero-posterior one. Cell movements in the proximal epiblast rapidly mirror this, so that cells expressing genes associated with formation of the primitive streak congregate posteriorly. Although Ding *et al.* found that antero-posterior patterning occurs in *Cripto* null embryos, it is not in the correct orientation (Fig. 1b) — the anterior tissue remains in a distal location,

whereas the most posterior epiblast derivative, extra-embryonic mesoderm, emerges proximally. Hence, the *Cripto* protein, secreted by epiblast, is necessary for the complementary cell movements in the visceral endoderm and epiblast that orientate the antero-posterior axis.

Other studies have revealed that extra-embryonic tissues are not simply nutritive — they also directly influence embryonic patterning. Removal of the anterior visceral endoderm inhibits differentiation of anterior neural tissue⁴. Expression of genes such as *nodal*⁵ (a member of the transforming growth factor- β (TGF- β) superfamily) and *Otx2* (ref. 6; a transcription factor) is required in the visceral endoderm, rather than the epiblast, for initiation of embryonic anterior pattern. Anterior development also needs Smad2 (a component of the TGF- β signal-transduction pathway) extra-embryonically^{7,8}. This all points to a mechanism for defining the anterior terminus that does not rely on signals from the classical embryonic organizer, but depends instead on interaction with the enveloping extra-embryonic tissues.

We do not yet know the details of this interaction, nor how the initial proximo-distal asymmetries arise. Likewise, the functions of *Cripto* and its relatives are not clear. We also need to know whether this extra-embryonic source of anterior patterning is a peculiarity of development in the uterus, or whether it has evolved from specialized cells with similar patterning functions in other vertebrate embryos. Where would such antecedents reside in lower vertebrates? In frogs, yolk cells in the deepest reaches of the organizer express *cerberus*, a gene that is instrumental in head formation⁹. Because a *cerberus* homologue is first expressed in the anterior visceral endoderm of mouse embryos¹⁰, the yolk cells may be the frog equivalent of extra-embryonic endoderm. In which case, do the same cues that induce the frog organizer induce these *cerberus*-expressing cells? Or, as in the mouse, does the tissue responsible for imparting anterior identity have a separate origin? □

Rosa Beddington is in the Division of Mammalian Development, MRC National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK.

e-mail: rbeddin@nimr.mrc.ac.uk

1. Harland, R. & Gerhart, J. *Annu. Rev. Cell Dev. Biol.* **13**, 611–667 (1997).
2. Ding, J. *et al. Nature* **395**, 702–707 (1998).
3. Thomas, P., Brown, A. & Beddington, R. S. P. *Development* **125**, 85–94 (1998).
4. Thomas, P. & Beddington, R. S. P. *Curr. Biol.* **6**, 1487–1496 (1996).
5. Varlet, I., Collignon, J. & Robertson, E. J. *Development* **124**, 1033–1044 (1997).
6. Rhinn, M., Dierich, A., LeMeur, M. & Ang, S. L. *Development* **125**, 845–856 (1998).
7. Waldrip, W. R. *et al. Cell* **92**, 797–808 (1998).
8. Nomura, M. & Li, E. *Nature* **393**, 786–790 (1998).
9. Bouwmeester, T. *et al. Nature* **382**, 595–601 (1996).
10. Belo, J. A. *et al. Mech. Dev.* **68**, 45–57 (1997).

Daedalus

New smells for old

Body odour is a worry for thousands of people. Chemical products alleged to abolish or modify it enjoy enormous sales. Daedalus now has a new approach

He points out that our body odour depends on our 'personal ecology' of skin bacteria, which make a living by fermenting our sweat. Some unfortunates are home to bacteria that ferment their sweat to highly offensive compounds. No amount of washing or deodorizing can abolish this problem. The skin bacteria form a stable equilibrium ecology, suited to the local conditions. They can never be completely eliminated, and the few survivors will recolonize their territory and establish the equilibrium anew.

But, says Daedalus, it is most unlikely that skin can sustain just one equilibrium ecology. He points out that our gut also harbours a complex stable consortium of organisms. Yet a few newcomers can sometimes precipitate an eruption of 'episodic gastroenteritis' that expels the old consortium and establishes a new one.

So DREADCO bacteriologists and cosmeticians are doing the same with the skin. They are sampling the skin flora of volunteers, and identifying the metabolic role of each organism in the consortium. For each volunteer, they will then devise a different mixture of organisms, calculated to exploit his sweat a little more efficiently, or be more at home in his surface microclimate. When spotted onto his skin, the new consortium will spread rapidly — possibly with a moving line of itching and reddening as the microscopic battle rages. When it has died down, his new occupiers will ferment his sweat differently. So he will smell different.

Discovering which consortia are most stable on what type of skin, and what smell they produce, will be a complex and demanding piece of research. But DREADCO's 'bioactive anti-odour treatment' should then be wildly popular. Victims of body odour will run (or be pushed) to the DREADCO clinic for the treatment. Criminals on the run may join the queue, hoping to change their smell and baffle the bloodhounds. And the Brussels bureaucracy, always keen to impose yet another intrusive European harmonization, may ask Daedalus to devise a super-dominant skin consortium, to be spotted onto every EU citizen. A uniform European smell, eliminating all xenophobic remarks about smelly foreigners, could bring a new instinctive unity to the emerging superstate.

David Jones

Obituary

Frederick Reines (1918–98)

Discoverer of the neutrino

For three months in 1951 a young Fred Reines sat in an empty office at Los Alamos, staring at a blank pad of paper on his desk. Following his final work on the effects of nuclear blasts at Eniwetok Atoll, he had requested a leave-in-residence to contemplate what he might do with his life. His years on the Manhattan Project under Hans Bethe and Richard Feynman placed him in the company of many of the greatest minds in science. The experience inspired him to seek some new way to address the central problems of physics.

His pondering spawned a bold idea. He would try to detect the free neutrino. Some 20 years before, Wolfgang Pauli had postulated the neutrino as a way of dealing with a disturbing violation of energy and momentum conservation observed in radioactive decay. Although he was not comfortable with his solution, Pauli proposed to save the conservation laws by assigning the missing energy and momentum to a ghostly new particle. As Pauli put it, "I have to replace something we do not understand with something else we cannot observe". Indeed, Bethe and Rudolf Peierls wrote in *Nature* in 1934, "...there is no practically possible way of observing the neutrino".

Yet, because the particle had later been embraced in a highly successful theory of β -decay by Enrico Fermi, at the time of Reines's deliberations nearly everyone believed that the neutrino was a reality. Why would Reines want to risk his career on a dubious attempt to prove what was already accepted? He claimed to be motivated simply by a stubborn desire to do the impossible. But his friends know better. Reines had an instinct — a feeling that there was something absolutely fundamental about the neutrino. In defiance of contemporary fashion, he set upon his heroically difficult quest.

Trained as a theorist, Reines had received his PhD from New York University in 1944. He would now have to become an experimentalist. He joined forces with Clyde Cowan, and together they contemplated a particle with no electric charge and little if any mass, that interacted so weakly with matter that the entire Earth is highly transparent to it. A fission bomb seemed the only adequate neutrino source; but fortunately they realized that, in a large liquid scintillation detector, the signature of an antineutrino captured on a proton would be so distinctive that the experiment could be performed under more controlled

conditions near a nuclear reactor.

The experiment was mounted at the Hanford plant in Washington state, where a hint of a signal was seen. Encouraged, they moved to a more favourable environment at a Savannah River reactor in South Carolina, and installed an improved detector. Following an exhaustive series of tests, the interaction of free antineutrinos on protons was firmly established in 1956. Reines and Cowan announced their victory in a telegram to Pauli at CERN, the European physics facility.

An entire new field of physics and astronomy had been opened up. Now that we could capture neutrinos from reactors, why not neutrinos from the Sun, from cosmic rays and from collapsing stars? Rightly known as the father of neutrino physics, it is Reines, perhaps more than any other scientist, who is identified with the discovery of a fundamental particle and its subsequent development as a tool for revealing new knowledge. He went on to exploit the huge antineutrino flux at Savannah River to establish basic properties of the neutrino and search for exotic new physics.

In 1959 he accepted a position as chair of physics at Case Institute of Technology, where he continued his pursuit of neutrino physics on many fronts. He built enormous scintillation and water Cerenkov detectors in deep mines to escape cosmic-ray background and to search for solar and extra-solar neutrinos, as well as neutrinos from the interaction of cosmic-ray primaries in the Earth's atmosphere. Another thread in his extraordinary career was his testing of conservation laws. No law was too sacred to escape his experimental scrutiny. He extended the limits on conservation of electric charge and of baryon number. He searched for lepton-number violation and evidence of neutrino-antineutrino identity in an experiment on neutrino-less double- β -decay, and he nurtured double- β -decay experiments done by others.

Among his 'firsts' were the detection of atmospheric neutrinos, of elastic scattering of antineutrinos on electrons, and of weak neutral current interactions of electron antineutrinos on deuterons; and the remarkable identification of a burst of neutrinos, from supernova 1987A, that confirmed the place of the neutrino in theories of stellar collapse.

Reines's philosophy of experimental physics is perhaps best illustrated by remarks he made while at the University of California at Irvine, where he had worked since his arrival as founding dean of

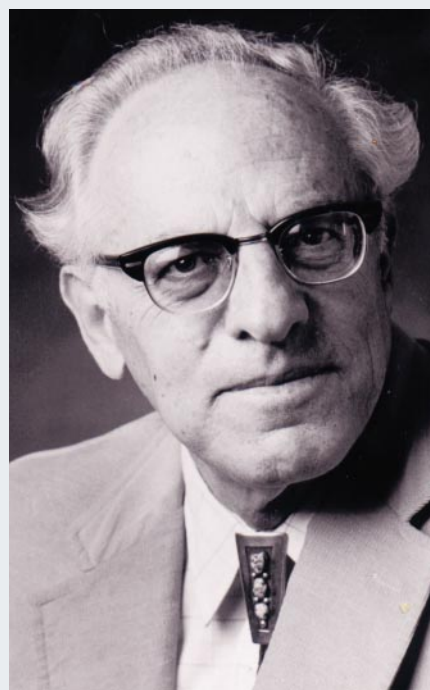
physical sciences in 1966. During the 20 years that the elastic scattering experiment was under way, guidance from theoretical developments was constantly changing. Had he been influenced by these developments it would have distracted him from the steadfastness that led to the eventual detection. In his own words, "This is not to say that experimentalists should proceed independently of theory, but it does suggest that the coupling should not be too tight".

A large man with a determined expression and a deep, booming voice, Reines was an imposing figure to his young graduate students. There was an intensity about him that told of a course well set, with no impending obstacle able to withstand his will. Yet beneath was a kind and playful spirit who whistled in the stair wells, sang songs from Gilbert and Sullivan, and who used his deep voice to great effect in telling jokes he had heard in Russia. While at Case, Fred contributed his fine baritone to the chorus of the Cleveland Symphony Orchestra.

For his discovery and pioneering investigation of the neutrino, Fred Reines shared the 1995 Nobel prize in physics with Martin Perl, discoverer of the tau lepton. Retired in 1988, he remained at Irvine as distinguished emeritus professor of physics until his death on 26 August. He will be missed by his many friends the world over.

Michael Moe

Michael Moe, 1 Mann Street, Irvine, California 92612-2747, USA, recently retired from the Department of Physics and Astronomy, University of California at Irvine.
e-mail: moe@alumni.stanford.org



UC, IRVINE

Calcium – a life and death signal

Michael J. Berridge, Martin D. Bootman & Peter Lipp

One of the most versatile and universal signalling agents in the human body is the calcium ion, Ca^{2+} . How does this simple ion act during cell birth, life and death, and how does it regulate so many different cellular processes?

Almost everything that we do is controlled by Ca^{2+} — how we move, how our hearts beat and how our brains process information and store memories. To do all of this, Ca^{2+} acts as an intracellular messenger, relaying information within cells to regulate their activity. For example, Ca^{2+} triggers life at fertilization, and controls the development and differentiation of cells into specialized types. It mediates the subsequent activity of these cells and, finally, is invariably involved in cell death. To coordinate all of these functions, Ca^{2+} signals need to be flexible yet precisely regulated. This incredible versatility arises through the use of a Ca^{2+} -signalling 'tool kit', whereby the ion can act in the various contexts of space, time and amplitude. Different cell types then select combinations of Ca^{2+} signals with the precise parameters to fit their physiology.

Space

At the cellular level, Ca^{2+} is derived from two sources — external and internal. It can enter from outside the cell by passing through channels that span the external

barrier, plasma membrane. Or it can be released from internal Ca^{2+} stores, through channels in the endoplasmic or sarcoplasmic reticulum^{1,2}, membranous networks that are also the site of protein synthesis and transport.

Improvements in imaging technology mean that cell physiologists can now see how the Ca^{2+} signals are generated. When a Ca^{2+} channel opens, a highly concentrated plume of Ca^{2+} forms around its mouth and then dissipates rapidly by diffusion after the channel has closed. Such localized signals, which can originate from channels in the plasma membrane or on the internal stores, represent the elementary events — the basic building blocks of Ca^{2+} signalling (Fig. 1a)^{3–5}. The spatio-temporal properties of these elementary events, such as Ca^{2+} sparks and Ca^{2+} puffs, differ depending on the nature and location of the channels. By characterizing these signals, we can discover how the Ca^{2+} -signalling repertoire is elaborated. Essentially, these elementary signals have two functions. They can either activate highly localized cellular processes in the immediate

vicinity of the channels (Fig. 1a) or, by recruiting channels throughout the cell, they can activate processes at a global level (Fig. 1b, c).

The subcellular location of Ca^{2+} channels is crucial for targeting elementary signals to different cellular processes. In smooth muscle, for example, Ca^{2+} sparks that arise locally, near the plasma membrane, activate potassium (K^+) channels (Fig. 1a), causing the muscle to relax^{6,7}. But when elementary-release events deeper in the cell are coordinated to create a global Ca^{2+} signal, the muscle contracts. This is a striking example of how spatial organization enables Ca^{2+} to activate opposing cellular responses in the same cell.

For sites of elementary Ca^{2+} release to produce global responses, the individual channels must communicate with each other, to set up Ca^{2+} waves (Fig. 1b). If cells are connected, such intracellular waves can spread into neighbouring cells and become intercellular waves to coordinate cellular responses within a tissue (Fig. 1c).

Time

One of the paradoxes surrounding Ca^{2+} is that it is a signal for both life and death — although elevations in Ca^{2+} are necessary for it to act as a signal, prolonged increases in the concentration of Ca^{2+} can be lethal. Cells avoid death either by using low-amplitude Ca^{2+} signals or, more usually, by delivering the signals as brief 'transients'. These principles apply to both elementary and global signals. Single transients are used to activate certain cellular processes, such as secretion of cellular material in membrane-

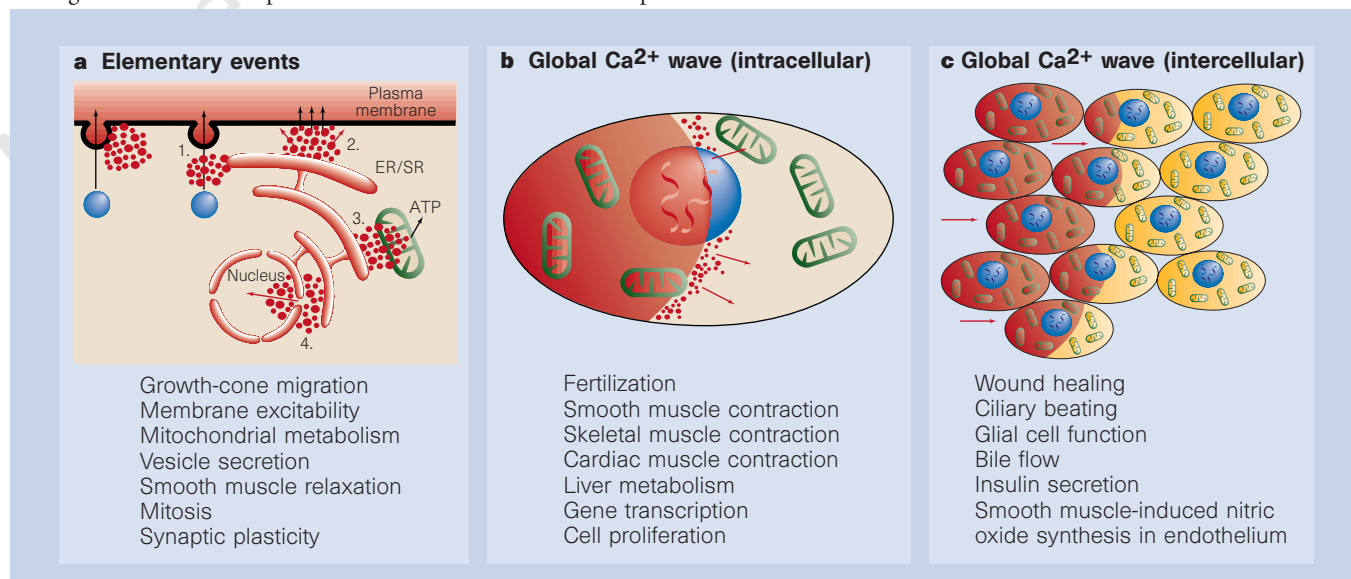


Figure 1 Spatial aspects of Ca^{2+} signalling. a, Elementary events (red) result from the entry of external Ca^{2+} across the plasma membrane or release from internal stores in the endoplasmic or sarcoplasmic reticulum (ER/SR). They generate localized concentrations of Ca^{2+} that can activate many processes, including export of cellular material (1), opening of K^+ channels (2) and metabolism in the mitochondria (3). The Ca^{2+} signals can also enter the nucleus (4). All of these processes respond to the very high concentrations of Ca^{2+} that build up within the sub-domain of the elementary events. b, Global Ca^{2+} signals are produced by coordinating the activity of elementary events to produce a Ca^{2+} wave that spreads throughout the cell. c, The activity of neighbouring cells within a tissue can be coordinated by an intercellular wave that spreads from one cell to the next.

bound vesicles, or muscle contraction. However, when information has to be relayed over longer time periods, cells use repetitive signals known as Ca^{2+} oscillations. Both the elementary events and the global signals can oscillate, but they have widely different periods. For example, whereas the period of elementary Ca^{2+} sparks in arterial smooth muscle is 0.1–0.5 seconds, it is 10–60 seconds for global waves in liver cells, 1–35 minutes for Ca^{2+} waves in human eggs after fertilization, and 10–20 hours for the spontaneous Ca^{2+} transients that control cell division.

Cells use frequency modulation (FM) to vary the intensity and nature of the physiological output. For instance, arteries can be made to dilate by increasing the frequency of Ca^{2+} sparks, which cause the smooth muscle lining the arteries to relax^{6,7}. And by varying the frequency of global Ca^{2+} signals, different genes can be activated⁸. To use FM signalling, cells have developed decoders that respond to the frequency and longevity of the Ca^{2+}

signals. Probably the best-known example is an enzyme called calmodulin-dependent protein kinase II, which is found in both animal and plant cells and which regulates other enzymes that rely on Ca^{2+} . It works by 'counting' Ca^{2+} transients⁹ and varying its activity accordingly. The enzyme is composed of many identical subunits, and these are activated to varying degrees depending on the frequency of the Ca^{2+} oscillations.

Amplitude

Information can also be encoded in the amplitude of Ca^{2+} signals. Such amplitude-modulated (AM) signalling is generally considered to be less reliable than that based on frequency, owing to the difficulties of detecting small Ca^{2+} changes above the background level. However, it has been shown that cells can interpret modest changes in the concentration of Ca^{2+} . For example, different genes can be activated by varying the amplitude of Ca^{2+} signals¹⁰.

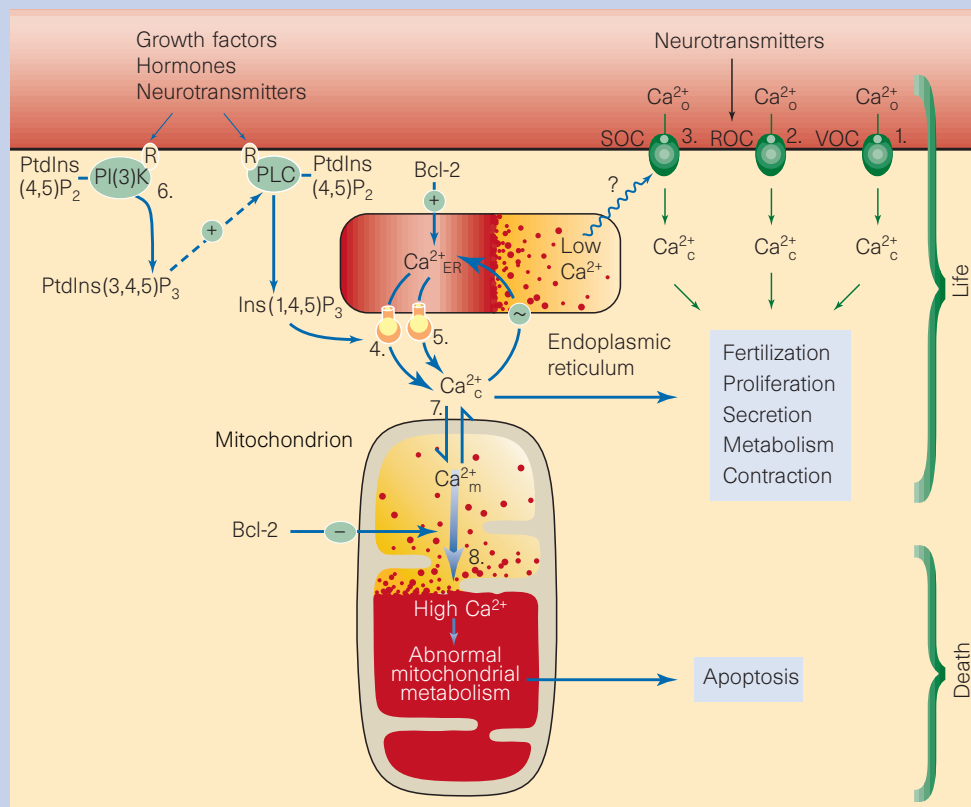
Fertilization and development

In mammals, life begins at fertilization when the sperm interacts with the egg to trigger a Ca^{2+} oscillation that persists for several hours. This prolonged period of repetitive Ca^{2+} pulses triggers the developmental programme by stimulating the enzymatic machinery involved in the cell-division cycle. There are no further changes in Ca^{2+} until the one-cell embryo is ready to divide, when a spontaneous Ca^{2+} transient triggers cleavage to form two daughter cells. There are indications that this orderly programme may be controlled by two distinct oscillators — Ca^{2+} signals and oscillating levels of proteins involved in the cell-division cycle. The Ca^{2+} oscillator seems to be the main mechanism, because it persists when dissociated from the cell-cycle oscillator¹¹. Just what drives this Ca^{2+} oscillator, with a period of 10–20 hours, is a mystery, but it may depend on periodic increases in the level of a diffusible intracellular messenger called inositol 1,4,5-trisphosphate ($\text{Ins}(1,4,5)\text{P}_3$)

Box 1 Basic mechanisms of Ca^{2+} signalling

Ca^{2+} signalling depends on increased levels of intracellular Ca^{2+} ($\text{Ca}^{2+}_{\text{c}}$), derived either from sources outside the cell ($\text{Ca}^{2+}_{\text{o}}$) or from stores within the endoplasmic reticulum ($\text{Ca}^{2+}_{\text{ER}}$). $\text{Ca}^{2+}_{\text{o}}$ may enter through (1) voltage-operated Ca^{2+} channels (VOCs) in excitable cells such as neurons or muscle cells, or (2) receptor-operated Ca^{2+} channels (ROCs) in response to neurotransmitters. Store-operated Ca^{2+} channels (SOCs; 3), which open when the internal stores are emptied of Ca^{2+} , are mainly found in non-excitable cells.

$\text{Ca}^{2+}_{\text{ER}}$ is released by two types of channel. Inositol 1,4,5-trisphosphate ($\text{Ins}(1,4,5)\text{P}_3$) is generated by the action of the enzyme phospholipase C (PLC) on phosphatidylinositol 4,5-bisphosphate ($\text{PtdIns}(4,5)\text{P}_2$) at the plasma membrane, in response to the action of growth factors, hormones or neurotransmitters at receptors (R). $\text{Ins}(1,4,5)\text{P}_3$ acts on receptors in the endoplasmic reticulum (4), which cause the release of $\text{Ca}^{2+}_{\text{ER}}$ from the store. Ryanodine receptors also



cause the release of $\text{Ca}^{2+}_{\text{ER}}$, especially in excitable cells (5).

In some cells, such as lymphocytes, the production of $\text{Ins}(1,4,5)\text{P}_3$ is modulated by the phosphatidylinositol 3-OH kinase, $\text{PI}(3)\text{K}$, signalling pathway, which uses

$\text{PtdIns}(4,5)\text{P}_2$ to produce the $\text{PtdIns}(3,4,5)\text{P}_3$ that acts as a messenger to maintain the activity of PLC. Some of the $\text{Ca}^{2+}_{\text{ER}}$ is rapidly taken up by the mitochondria ($\text{Ca}^{2+}_{\text{m}}$) and is then returned to the endoplasmic reticulum (7), although most of the stored

$\text{Ca}^{2+}_{\text{ER}}$ resides in the lumen of the endoplasmic reticulum. But if the mitochondria become overloaded with $\text{Ca}^{2+}_{\text{m}}$, the result is abnormal mitochondrial metabolism (8), which may activate programmed cell death. **M.J.B., M.D.B. & P.L.**

with each division¹². In many cells, hormones and growth factors activate the enzyme phospholipase C, which catalyses the production of $\text{Ins}(1,4,5)\text{P}_3$ from the membrane lipid phosphatidylinositol 4,5-bisphosphate (Box 1).

As embryos grow and groups of cells differentiate to perform specialized functions, Ca^{2+} signalling contributes to body polarity and pattern formation. For instance, in amphibians and zebrafish it is thought to help specify which cells will form structures at the top (dorsal) or the bottom (ventral) part of the embryo. When this dorso-ventral axis is being specified, there is a spontaneous increase in the level of $\text{Ins}(1,4,5)\text{P}_3$, possibly in the form of a gradient ranging from low levels on the dorsal side to high levels on the ventral side¹³. Indeed, if an antibody is added to prevent the cells from detecting $\text{Ins}(1,4,5)\text{P}_3$, ventral cells are converted to dorsal cells¹⁴. The increase in $\text{Ins}(1,4,5)\text{P}_3$ in zebrafish at least, coincides with the appearance of Ca^{2+} spikes restricted to a small group of cells. These spikes may be involved in dorso-ventral specification, because their frequencies are sensitive to the expression of developmental genes¹⁵.

Later in development, when the final form of the embryo is emerging, Ca^{2+} signals are again used to control the differentiation of specific cell types. For example, spontaneous Ca^{2+} transients control assembly of the contractile machinery during the formation of muscle¹⁶. In development of the heart, the Ca^{2+} -dependent 'nuclear factor of activated T cells' (NF-AT) helps to form the cardiac septum and valves¹⁷. Ca^{2+} is also involved in development of the nervous system. Spontaneous Ca^{2+} transients with different frequencies are responsible for various aspects of differentiation in developing amphibian neurons¹⁸. In the main body of the neuron, the natural frequency of two to four Ca^{2+} transients per hour is optimal for maturation of K^+ channels. But growth of projections called neurites, out in the extremities of the neuron, is tuned to higher Ca^{2+} signal frequencies. Ca^{2+} signals also help to control the migration of neuronal cells¹⁹, and the initial wiring-up processes that produce the complex neuronal circuits of the developing brain²⁰.

Cell proliferation

A prolonged period of Ca^{2+} signalling — similar to that in fertilization — is an important growth signal for many cells²¹. Alterations in Ca^{2+} signalling can underlie defects in cell growth and are implicated in some cancers. Conversely, cell proliferation can be reduced by interfering with the generation or action of Ca^{2+} signals. Because internal Ca^{2+} stores are finite, prolonged bouts of signalling depend on the influx of external Ca^{2+} through so-called store-operated Ca^{2+} channels (SOCs) in the plasma membrane

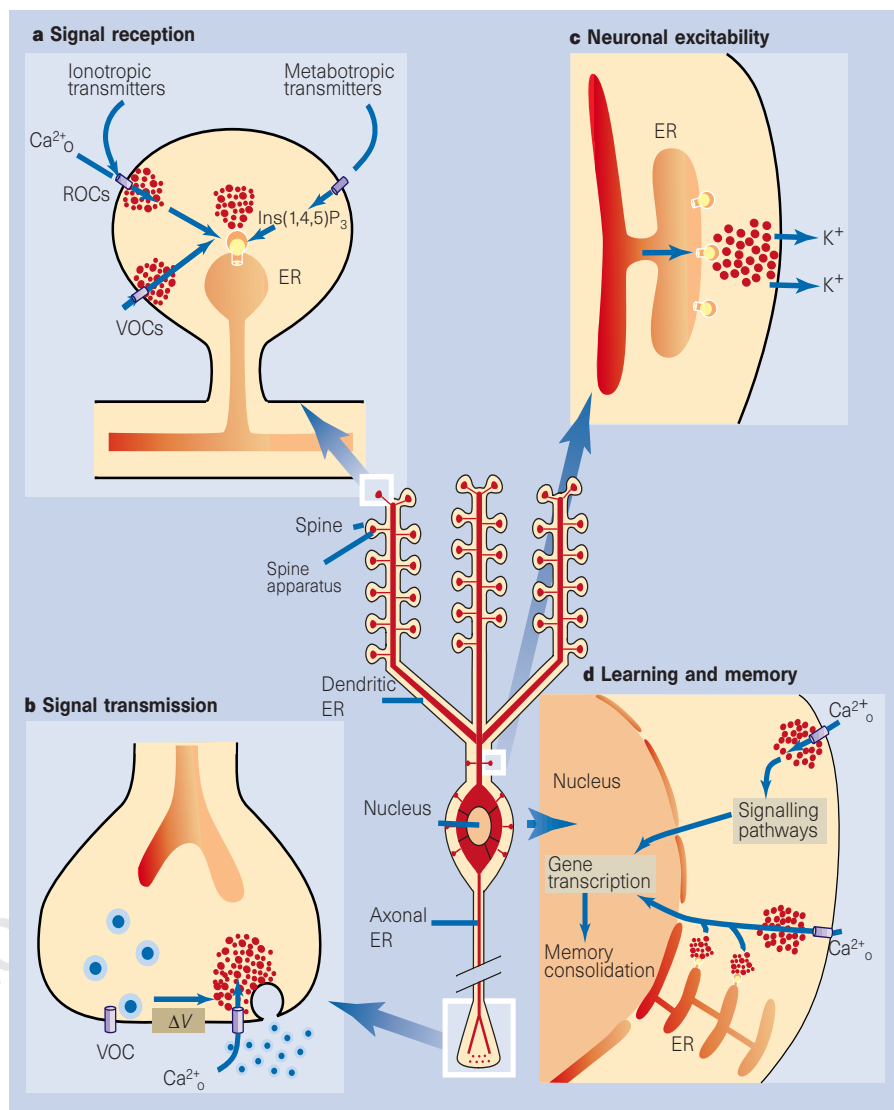


Figure 2 Compartmentalization of Ca^{2+} signals in neurons. **a**, External Ca^{2+} enters through receptor-operated channels (ROCs) or voltage-operated channels (VOCs), and the signal may be amplified by activating receptors for $\text{Ins}(1,4,5)\text{P}_3$ on the spine apparatus. By integrating separate input signals, these receptors could act as coincidence detectors. **b**, In response to electrical signals (ΔV), the entry of Ca^{2+} through VOCs stimulates neurotransmitter release. Such signals, arising through Ca^{2+} entry, are usually very rapid in onset and decline. **c**, Localized Ca^{2+} release from the endoplasmic reticulum (ER) opens K^+ channels to modulate neuronal excitability. **d**, Proposed role of Ca^{2+} in memory consolidation. Entry of external Ca^{2+} can act locally by co-opting other signalling pathways, or it can act globally by flooding directly into the nucleus. Such a global signal might be amplified by release of Ca^{2+} from the internal stores.

(Box 1). These SOC might be the target of a Ca^{2+} -influx inhibitor, which has been found, in clinical trials, to slow down the growth of certain aggressive cancer cells²².

Ca^{2+} is also involved in the proliferation of immune cells (lymphocytes) in response to foreign antigens. Both T and B lymphocytes detect antigens through complex receptors on their surface. When a foreign molecule binds to an antigen receptor, $\text{Ins}(1,4,5)\text{P}_3$ is produced, stimulating the release of Ca^{2+} from internal stores. Once these stores are empty, entry of external Ca^{2+} is activated through the SOC (Box 1)²¹, allowing lymphocytes to maintain a prolonged increase of Ca^{2+} . This increase, which

often occurs as a series of regular Ca^{2+} oscillations, activates factors such as NF-AT, which enter the nucleus and cause genes to be turned on. The importance of Ca^{2+} signalling in lymphocyte activation is highlighted by the fact that immunosuppressant drugs, such as cyclosporin, work by inhibiting the Ca^{2+} -dependent activation of NF-AT.

Neurons

Neurons provide an excellent example of how a combination of elementary and global Ca^{2+} signals have been adapted to regulate a range of processes in a single cell²³. For example, Ca^{2+} is pivotal in receiving and transmitting neuronal signals, as well as in regulating

excitability and the changes that underlie learning and memory (Fig. 2). Ca^{2+} influx into the neuron has long been regarded as the main source of Ca^{2+} signals, but the significance of Ca^{2+} release from internal stores such as the endoplasmic reticulum is now becoming apparent. This continuous network extends throughout the neuron, and can be considered as a neuron within a neuron — it has properties that resemble those of the plasma membrane²³. There are many examples of how the plasma membrane and endoplasmic reticulum interact to produce spatially restricted signals in some regions, whereas in others they create global Ca^{2+} signals over large parts of the neuron.

Neurons have four main regions — a dendritic tree containing the spines that receive inputs from other neurons, a cell body that contains the nucleus, a long axon that conveys electrical signals, and the synaptic ending, which transmits these signals to target cells such as other neurons, muscle cells or gland cells (Fig. 2). The neuronal spines act as microprocessors within the brain and, although they have a volume of only around $0.1 \mu\text{m}^3$, they can function as autonomous compartments. There is increasing evidence that Ca^{2+} signals within spines may be responsible for the 'synaptic plasticity' that leads to short-term memory.

When a signal is received, Ca^{2+} either enters the spines from the outside or is released from the spine apparatus, a continuation of the endoplasmic reticulum that bears channels such as those activated by $\text{Ins}(1,4,5)\text{P}_3$. Because these receptors can be stimulated by both $\text{Ins}(1,4,5)\text{P}_3$ and Ca^{2+} , they could act as coincidence detectors to integrate information from separate neural signals (Fig. 2a)²³. For example, signals that promote the entry of Ca^{2+} will cooperate with those that generate $\text{Ins}(1,4,5)\text{P}_3$. Such concurrent stimuli will promote the explosive release of stored Ca^{2+} , bringing about the modifications that are thought to be responsible for learning and memory.

One puzzle is that Ca^{2+} has been implicated in both stimulating and depressing the transmission of nervous signals. Subtle modifications in the amplitude, or spatial and temporal presentation of Ca^{2+} signals, as described earlier, may account for this. Spatially restricted Ca^{2+} signals also control neuronal excitability (Fig. 2c). When neurons fire, information is relayed down the axon to the synaptic ending, activating the secretion of neurotransmitters — chemicals that excite neighbouring neurons. Elementary Ca^{2+} signals are used to produce brief, highly localized transients that trigger release of vesicles containing the neurotransmitter (Fig. 2b)²⁴. In the cell body itself, elementary Ca^{2+} signals can modulate neuronal excitability by activating Ca^{2+} -dependent K^+

channels. These channels allow an efflux of K^+ ions through the plasma membrane, inhibiting subsequent electrical activity (Fig. 2c).

To create more permanent memories, the short-term modifications described above have to be consolidated by information from the nucleus. Once again, Ca^{2+} is involved (Fig. 2d)^{25,26}. But how can the information arriving at the synaptic ending trigger gene transcription far away in the nucleus? To do this, Ca^{2+} seems to recruit additional signalling components that migrate into the nucleus and activate genes there²⁵. In addition, the Ca^{2+} signals themselves, derived from either the entry or release of Ca^{2+} , can also activate genes in the nucleus²⁶.

Cell death

Very high concentrations of Ca^{2+} can lead to the disintegration of cells (necrosis) through the activity of Ca^{2+} -sensitive protein-digesting enzymes. Calcium has also been implicated in the more orderly programme of cell death known as apoptosis. Apoptosis is important during both normal development (the formation of tissue patterns, for example) and pathological conditions such as AIDS, Alzheimer's disease and cancer. A protein that is mutated in cancerous cells, called Bcl-2, prevents the cell death that would normally limit the survival and proliferation of cancer cells. Bcl-2 mediates some of its anti-apoptotic action by modifying the way in which organelles such as the endoplasmic reticulum and mitochondria (where respiration occurs) handle Ca^{2+} (Box 1).

But what is the relationship between Bcl-2, Ca^{2+} signalling and apoptosis? In many cells, mitochondria participate in the recovery phase of normal Ca^{2+} transients — they sequester some of the Ca^{2+} signal, which is later returned to the endoplasmic reticulum. So, during normal Ca^{2+} signalling, there is a continuous shuttling of Ca^{2+} between these two organelles. Normally, most of the Ca^{2+} resides within the lumen of the endoplasmic reticulum, with very little in the mitochondria. These high levels of Ca^{2+} are essential. Not only do they form a reservoir of signal Ca^{2+} in the endoplasmic reticulum, but they are also essential for the synthesis and processing of proteins there.

If the Ca^{2+} stored within the endoplasmic reticulum was depleted, the mitochondria would become overloaded and there would be two main consequences for the cells. First, the decline in levels of Ca^{2+} in the endoplasmic reticulum would lead to the activation of stress signals, which switch on the genes associated with cell death. Interestingly, some of these genes also specify proteins that bind Ca^{2+} in the endoplasmic reticulum, and this could be a desperate attempt by the cell to restore the correct balance of Ca^{2+} between the endoplasmic reticulum and the mitochondria. Second, the build-up of mito-

chondrial Ca^{2+} initiates a programme of events that leads to cell death (Box 1). In normal cells, Bcl-2 may modify the Ca^{2+} -handling properties of the endoplasmic reticulum²⁷ and the mitochondria²⁸ (Box 1), to restore the correct Ca^{2+} balance.

Conclusion

This brief — and all too simplistic — journey through some of the processes controlled by Ca^{2+} illustrates the universality of this signalling mechanism, which triggers a new life at fertilization and is then re-used over and over again to regulate the developmental programme as it unfolds to produce a new organism. As cells differentiate to perform different functions, they select out those components of the Ca^{2+} -signalling toolkit that best fit their remit. This versatility is emphasized by growing evidence that Ca^{2+} controls cellular processes as diverse as cell proliferation and the neuronal plasticity that is responsible for learning and memory. But at any moment, any of these orderly signalling events can be switched to activate a programme that leads to cell death — a big challenge for the future is to understand how Ca^{2+} suddenly transforms from a signal for life to a signal of death. □

Michael J. Berridge, Martin D. Bootman and Peter Lipp are in the Laboratory of Molecular Signalling, The Babraham Institute, Babraham, Cambridge CB2 4AT, UK.

e-mail: michael.berridge@bbsrc.ac.uk

- Berridge, M. J. *Nature* **361**, 315–325 (1993).
- Clapham, D. E. *Cell* **80**, 259–268 (1995).
- Bootman, M. D. & Berridge, M. J. *Cell* **83**, 675–678 (1995).
- Lipp, P. & Niggli, E. *Prog. Biophys. Mol. Biol.* **65**, 265–296 (1996).
- Berridge, M. J. *J. Physiol., Lond.* **499**, 291–306 (1997).
- Nelson, M. T. *et al. Science* **270**, 633–637 (1995).
- Porter, V. A. *et al. Am. J. Physiol.* **274**, C1346–C1355 (1998).
- Dolmetsch, R. E., Xu, K. & Lewis, R. S. *Nature* **392**, 933–936 (1998).
- De Koninck, P. & Schulman, H. *Science* **279**, 227–230 (1998).
- Dolmetsch, R. E., Lewis, R. S., Goodnow, C. C. & Healy, J. I. *Nature* **386**, 855–858 (1997).
- Swanson, C. A., Arkin, A. P. & Ross, J. *Proc. Natl Acad. Sci. USA* **94**, 1194–1199 (1997).
- Ciapa, B., Pesando, D., Wilding, M. & Whitaker, M. *Nature* **368**, 875–878 (1994).
- Ault, K. T., Durmowicz, G., Galione, A., Harger, P. L. & Busa, W. B. *Development* **122**, 2033–2041 (1996).
- Kume, S. *et al. Science* **278**, 1940–1943 (1997).
- Stusarski, D. C., Corcos, V. G. & Moon, R. T. *Nature* **390**, 410–413 (1997).
- Ferrari, M. B., Rohrbough, J. & Spitzer, N. C. *Dev. Biol.* **178**, 484–497 (1996).
- de la Pompa, J. L. *et al. Nature* **392**, 182–185 (1998).
- Gu, X. & Spitzer, N. C. *Nature* **375**, 784–787 (1995).
- Komura, H. & Rakic, P. *Neuron* **17**, 275–285 (1996).
- Katz, L. C. & Shatz, C. J. *Science* **274**, 1133–1138 (1996).
- Berridge, M. J. *BioEssays* **17**, 491–500 (1995).
- Kohn, E. C. *et al. Cancer Res.* **56**, 569–573 (1996).
- Berridge, M. J. *Neuron* **21**, 13–26 (1998).
- Sugimori, M., Lang, E. J., Silver, R. B. & Llinas, R. *Biol. Bull.* **187**, 300–303 (1994).
- Bitto, H., Deisseroth, K. & Tsien, R. W. *Curr. Opin. Neurobiol.* **7**, 419–429 (1997).
- Hardingham, G. E., Chawla, S., Johnson, C. M. & Bading, H. *Nature* **385**, 260–265 (1997).
- He, H., Lam, M., McCormick, T. S. & Distelhorst, C. W. *J. Cell Biol.* **138**, 1219–1228 (1997).
- Murphy, A. N., Bredesen, D. E., Cortopassi, G., Wang, E. & Fiskum, G. *Proc. Natl Acad. Sci. USA* **93**, 9893–9898 (1996).

Monge's maths, Hummel's highlights

The type of perspective favoured in technical drawing was first used by military draughtsmen to ensure accurate fortifications. It was developed geometrically by Gaspard Monge and drawn to perfection by Johann Hummel.

Martin Kemp

Linear perspective is about the portrayal of the position of objects in space. Yet in its pictorial mode, it creates an illusion which can only be used to provide measured definition of the dimensions of forms if the diminutions are reversed through a knowledge of the parameters of the construction. It was the special mode developed for technical drawing, above all for fortifications, that both presented spatial effects and retained dimensional accuracy. It became the speciality of military draughtsmen, engineers and stone-cutters in the eighteenth century.

What was known as *perspectiva militaris*, and is now called orthographic or isometric perspective, set the vanishing point at a notionally infinite distance, in such a way that parallels do not converge, and distant forms can be measured on the same scale as adjacent ones.

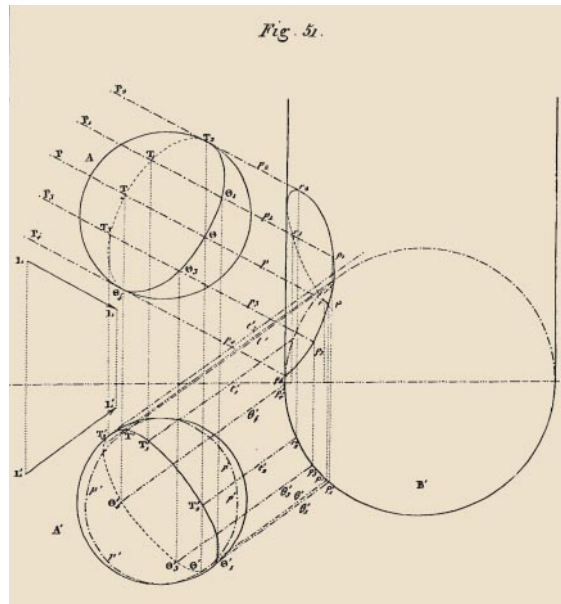
Gaspard Monge, the founder of descriptive geometry, was educated in this tradition at the French military academy at Mézières. During the Revolutionary era, he reshaped mathematics teaching in the newly reformed school system.

His *Descriptive Geometry* of 1799, edited from his lectures at the Ecole Normale, zealously advocated his method "to define the position of a point in space" by reference to "those other objects that are of known position in some distinctive part of space", most directly by two intersecting planes. His procedures, closely related to shadow projection techniques in the more technical books on artists' perspective, revived the power of drawn geometrical procedures after years of dominance by algebraic analysis.

Monge himself extolled descriptive geometry as "a means of investigating the truth [that] should necessarily be introduced into the plan of national education", and as a prime tool for the artist, for whom "nothing is arbitrary" with respect to formal definition – even when complex, curved surfaces are involved.

His methods had a huge impact on technical drawing in Europe and America, but few artists were predisposed to follow the demanding letter of his law.

A striking exception was the professor of optics at the Berlin Royal Academy of Art, Johann Hummel. A rather graceless draughtsman of figures, Hummel excelled in perspectival views. His *pièce de résistance* is a series of paintings on the making and erec-



"Projection of a shadow cast by a sphere on a cylindrical surface", from Monge's *Descriptive Geometry*, 1799.



Hummel's *Polishing of the Granite Bowl*, 1831, Berlin.

tion of the huge granite bowl installed on the Lustgarten in Berlin in 1831. The complex curvature of the bowl in itself set a stiff task for any conscientious perspectivist, but the geometry of the reflections presented problems that only the most dedicated and well-informed investigator would attempt to resolve.

Well aware of the most technical accounts of geometrical projection, particularly of shadows on convex and concave surfaces of complex curvature, he used his pictures of the bowl, most strikingly in its inverted position while being polished, as virtuoso demonstrations of the intriguing configura-

tions of geometrical forms 'morphed' by reflection.

As Jonathan Miller shows in "Mirror Image", his fascinating exhibition at the National Gallery in London (on until 13 December), the artist relies upon our perceptual ability to read the shapes as highlights through our grasp of context. Hummel moulds the reflected contours of the leaded windows like a musician shaping a phrase. The result is both precise and dazzling. □

Martin Kemp is in the Department of the History of Art, University of Oxford, 35 Beaumont Street, Oxford OX1 2PG, UK.

e-mail: martin.kemp@trinity.oxford.ac.uk

Female gene flow stratifies Hindu castes

Scientists have long been interested in understanding how social processes modulate evolutionary forces¹. A good example of this is the intensively studied Hindu caste system, which governs the mating practices of nearly one-sixth of the world's population^{2,3}. However, there is controversy concerning the effect of social stratification on the genetic structure of caste communities. Here we show that differences in social rank between castes correspond to mitochondrial DNA (mtDNA) distances between castes but not genetic distances estimated from Y-chromosome data.

Hindu populations are stratified into five *varna*, each of which is associated with a 'status'. Ranked from lowest to highest status, these are Panchama, Sudra, Vysya, Kshatriya and Brahmin⁴. There are approximately 2,000 castes in contemporary India, roughly grouped into these *varna*. Religious rites, access to material resources and education, and occupational specialization are directly related to caste status. Marriages between individuals of equal status are preferred. Matings between a man from a higher *varna* and a woman from a lower *varna* are permissible under certain circumstances, in which case the offspring tend to attain a status similar to that of their father. In contrast, marriage of a woman from a higher *varna* to a man of a lower *varna* is strongly discouraged. This suggests that women have limited but upward social mobility, whereas men have very little⁴.

To test the effects of this mating system on genetic variation, we analysed mtDNA and Y-chromosome variation in 250 individuals from 12 Telugu-speaking caste populations from northeastern Andhra Pradesh in southern India. Castes and *varna* were ranked by status into upper (Brahmin, Kshatriya and Vysya), middle (Kapu, Yadava, Jalari and Wadabahija), and lower (Relli, Madiga and Mala) groups.

We identified 182 unique mtDNA haplotypes, four of which are shared among upper, middle and lower castes, seven between lower and middle castes, and three between middle and upper castes. No haplotypes are shared exclusively between upper and lower castes. This suggests that haplotype sharing between castes is limited by social rank. The genetic distance between upper and lower castes (0.00045) is 1.5 times greater than between the upper and middle (0.00024) or middle and lower castes (0.00030). Thus, mtDNA distances between castes of different status are stratified according to social rank (Fig. 1a). This is consistent with higher levels of female gene flow between more closely ranked castes.

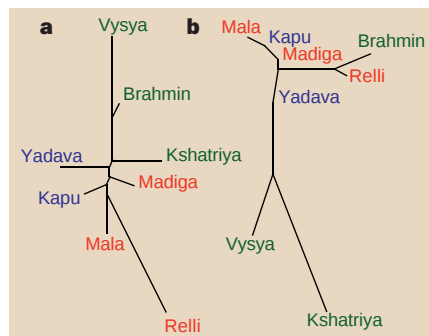


Figure 1 Neighbour-joining networks⁷ of genetic distances among caste communities. **a**, Network estimated from 411 base pairs of hypervariable region 1 of mtDNA⁸. Distances between upper castes (Brahmin, $n = 41$; Vysya, $n = 10$; and Kshatriya, $n = 10$), middle castes (Yadava, $n = 48$; and Kapu, $n = 52$), and lower castes (Relli, $n = 20$; Mala, $n = 25$; and Madiga, $n = 28$) are correlated with social rank. Pairwise estimates of the proportion of genetic variation attributable to differences between groups indicate that the genetic distance between upper and lower castes is fourfold higher than between upper and middle castes ($P < 0.001$). A neighbour-joining network of unique haplotypes reveals a star-like pattern (data not shown) with a few short central branches linking nodes from which tufts of long branches emerge. Lineages from different castes are scattered throughout the network, consistent with gene flow between communities or the sorting of lineages before the separation of castes. Analyses limited to transversions or excluding rapidly mutating sites within mtDNA HVS1 (ref. 9) do not substantially alter these results (data not shown). **b**, Network of genetic distances among caste communities estimated from Y-chromosome STRs (DYS19, DYS288, DYS388, DYS389A, DYS389B, DYS390 and DYS392) and SNPs (SRY10831, SY57 and RPS4Y)⁹. Upper, middle and lower castes do not cluster together, and distances between upper, middle and lower castes are not correlated with social rank.

Analysis of Y-chromosome data revealed that seven short-tandem repeats (STRs) and three single-nucleotide polymorphisms (SNPs) produced 107 unique haplotypes in 194 individuals for whom the data were complete. For the STR data, the genetic distance between upper and lower castes (0.0054) is similar to that between the upper and middle castes (0.0062), and larger than that between middle and lower castes (0.0005). Thus, caste status is not associated with Y-chromosome genetic distance (Fig. 1b).

A statistical comparison of the genetic distances based on mtDNA and Y-chromosome data reveals a low, non-significant correlation ($r = 0.14$, $P > 0.3$) by the Mantel permutation test⁵. There are several possible explanations for this lack of concordance.

First, there may not be enough Y-chromosome STRs to reveal a meaningful pattern, although a comparison of a distance matrix based on the Y-chromosome SNPs with the matrix based on STRs yields a significant positive correlation ($r = 0.68$, $P < 0.0025$). This high degree of consistency would not be expected if the STR distances were due to insufficient data.

Second, the lack of correlation between social rank and Y-chromosome distance might be produced by random male movement between castes. This explanation is unlikely because the Y-chromosome data show greater differentiation between groups than the mtDNA data: the proportion of genetic variation attributable to differences between groups for Y-chromosome STRs is 0.10 ($P < 0.001$), whereas that based on mtDNA data is only 0.015 ($P < 0.002$). This result is consistent with females having greater intercaste gene flow than males, and agrees with historical evidence that males were less likely to change castes.

A third possible explanation is that the discordance is caused by differences in mutation rates between the Y-chromosome and mtDNA systems⁶. However, the greatest observed difference in mutation rates is between the Y-chromosome STRs and the Y-chromosome SNPs, yet these two data sets are highly concordant.

The remaining explanation for the discordance, and the one most consistent with the historical data, is that systematic female gene flow among castes produced a correlation between social rank and mtDNA distances. A relative lack of male gene flow resulted in no correlation between social rank and Y-chromosome distances, and supports the idea that the pattern of Y-chromosome variation is largely the result of mutation and genetic drift. This indicates that mobility between castes of different status has been higher for females than males, as predicted from ethnographic records. We therefore conclude that genetic stratification of the Hindu caste system is driven by the social mobility of women.

Michael J. Bamshad*, W. Scott Watkins†, Mary E. Dixon‡, Lynn B. Jorde‡, B. Bhaskara Rao‡, J. M. Naidu‡, B. V. Ravi Prasad‡, Arani Rasanayagam§, Mike F. Hammer§

Departments of *Pediatrics and †Human Genetics, Eccles Institute of Human Genetics, University of Utah, Salt Lake City, Utah 84112, USA
e-mail: mike@thor.med.utah.edu

‡Department of Anthropology, Andhra University, Visakhapatnam, Andhra Pradesh, India

§Laboratory of Molecular Systematics and

Evolution, University of Arizona,
Tucson, Arizona 85721, USA

- Jorde, L. B., Bamshad, M. & Rogers, A. R. *BioEssays* **20**, 126–136 (1998).
- Majumder, P. P. *Evol. Anthropol.* **6**, 100–110 (1998).
- Bamshad *et al.* *Hum. Biol.* **68**, 1–33 (1996).
- Tambia, S. J. in *The Character of Kinship* (ed. Goody, J.) 191–229 (Cambridge Univ. Press, 1973).
- Mantel, N. *Cancer Res.* **27**, 209–220 (1967).
- Heyer, E. *et al.* *Hum. Mol. Genet.* **6**, 799–803 (1997).
- Saitou, N. & Nei, M. *Mol. Biol. Evol.* **4**, 406–425 (1988).
- Anderson *et al.* *Nature* **290**, 457–465 (1981).
- Wakeley, J. J. *Mol. Biol. Evol.* **37**, 613–623 (1993).
- Hammer, M. F. *et al.* *Mol. Biol. Evol.* **15**, 427–441 (1998).

Effects of progesterone or neuroactive steroid?

Smith *et al.*¹ reported some interesting results about the effects of progesterone treatment and withdrawal on the gene encoding the GABA_A receptor $\alpha 4$ subunit, a constituent of receptors for the neurotransmitter GABA (γ -aminobutyric acid). However, we disagree with their conclusion that the effects of progesterone withdrawal are mediated by reduced levels in the brain of the GABA_A receptor active neurosteroid 3 α -OH-5 α -pregnan-20-one (3 α ,5 α -THP, or allopregnanolone). The ability of indomethacin to reverse the effects of progesterone was interpreted as evidence that the effects of progesterone were mediated by 3 α ,5 α -THP. Smith *et al.*¹ did not provide direct evidence that levels of 3 α ,5 α -THP were reduced by indomethacin (which reversed the effects of progesterone), or that progesterone levels were not altered by indomethacin. In some cases, indomethacin can increase progesterone levels directly², which may explain why indomethacin can reverse the effects of progesterone withdrawal.

The effect of indomethacin blockade of 3 α -hydroxysteroid oxidoreductase activity is dependent on the relative levels of 3 α ,5 α -THP and its immediate precursor 5 α -dihydroprogesterone. When 3 α ,5 α -THP concentrations are raised after progesterone infusion or stress, indomethacin would block the oxidation of 3 α ,5 α -THP, increasing its concentration. Using the same dose regimen as ref. 1 we find that stress-induced elevation of 3 α ,5 α -THP levels in the brain are increased by pretreatment with indomethacin (Fig. 1). Blockade of this enzyme with fluoxetine has also been shown to raise 3 α ,5 α -THP levels in the rat brain³. The omission of 3 α ,5 α -THP and progesterone levels in the report by Smith *et al.*¹ therefore raises concerns about the conclusion that 3 α ,5 α -THP mediates the effects of progesterone withdrawal.

Manipulation of progesterone levels in female rats is likely to influence the expression of many genes as well as the levels of

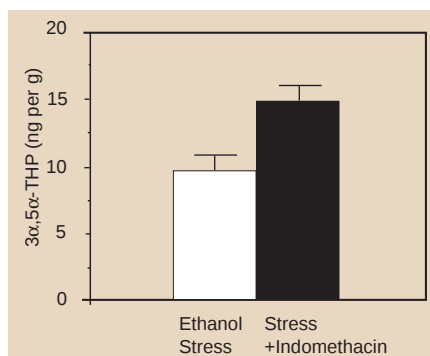


Figure 1 Indomethacin increases stress-induced 3 α ,5 α -THP levels in rat cerebral cortex. Male Sprague-Dawley rats were habituated to handling and injection. Indomethacin (0.1 mg per kg, intraperitoneal) was administered 20 minutes before ethanol injection stress (2 g per kg, intraperitoneal) and 3 α ,5 α -THP levels were measured as described⁴. Indomethacin increased stress-induced 3 α ,5 α -THP levels by $58.5 \pm 62.0\%$ ($P < 0.01$, Student's *t*-test).

numerous steroids other than 3 α ,5 α -THP. Therefore, even if 3 α ,5 α -THP levels are reduced by indomethacin, this does not demonstrate that the effects of progesterone withdrawal on GABA_A receptors are mediated by 3 α ,5 α -THP. Correlational evidence, although suggestive, does not establish a causal relationship.

The effects of progesterone and ethanol withdrawal on GABA_A receptor $\alpha 4$ subunit gene expression are similar. However, cross-tolerance between ethanol and 3 α ,5 α -THP does not occur *in vivo*. Indeed, chronic ethanol administration produces cross-tolerance to benzodiazepines, but also results in sensitization to both behavioural and neurochemical responses to 3 α ,5 α -THP⁴. Moreover, this effect is greater in female than male rats, suggesting that the higher progesterone levels in female rats are associated with greater sensitization to 3 α ,5 α -THP⁵. This sensitization is probably due to intrinsic changes in GABA_A receptors, as concentrations of 3 α ,5 α -THP are not altered after chronic ethanol administration⁴. Although the effects of ethanol^{4,6} and progesterone on GABA_A receptor $\alpha 4$ gene expression are known, there is no direct evidence that 3 α ,5 α -THP alters the expression of this gene.

A. Leslie Morrow*, **Margaret J. VanDoren**,
Leslie L. Devaud

Departments of Psychiatry and *Pharmacology,
and Center for Alcohol Studies,
University of North Carolina School of Medicine,
Chapel Hill, North Carolina 27599-7178, USA
e-mail: morrow@med.unc.edu

Smith *et al.* reply — Morrow *et al.* are incorrect to state that we did not provide direct evidence that 3 α -OH-5 α -pregnan-20-one (3 α ,5 α -THP, or allopregnanolone) levels were reduced by indomethacin. We pointed out¹ that our previous studies⁷

showed that indomethacin administration concomitant with progesterone reduced the level of 3 α ,5 α -THP in the central nervous system from 8.2 ± 1.3 ng per g to 2.3 ± 0.35 ng per g without altering levels of 5 α -DHP or progesterone. This treatment, which prevented the withdrawal of 3 α ,5 α -THP that normally accompanies progesterone withdrawal, completely prevented the withdrawal effects of progesterone we observed¹, including acceleration in the decay of GABA-gated current and upregulation of the GABA_A receptor $\alpha 4$ subunit.

We have shown⁸ that direct withdrawal of 3 α ,5 α -THP, in the absence of progesterone withdrawal, also results in faster decay of GABA-gated current in association with upregulation of the $\alpha 4$ subunit. In this case, withdrawal from the GABA-modulatory metabolite 3 α ,5 α -THP was accomplished with the use of a 5 α -reductase inhibitor to block 3 α ,5 α -THP formation without altering progesterone levels. Taken together, these results establish 3 α ,5 α -THP as the active agent that exhibits withdrawal properties. In contrast, Morrow *et al.* administer indomethacin 20 minutes before ethanol stress, which would prevent direct conversion of 3 α -DHP to 3 α ,5 α -THP. The indomethacin-induced increases in 3 α ,5 α -THP levels they report after ethanol stress can be explained only by stress-induced effects on the degradation of 3 α ,5 α -THP.

We further disagree with the suggestion by Morrow *et al.* that our findings are purely correlational. Indomethacin is a highly selective blocker of 3 α -hydroxysteroid oxidoreductase activity⁹, and so reduces 3 α ,5 α -THP levels specifically, without altering levels of other hormones such as 5 α -DHP or progesterone. This is therefore the most definitive procedure to determine the role of the 3 α ,5 α -THP metabolite in the upregulation of the $\alpha 4$ subunit we observe after progesterone withdrawal. Direct administration of 3 α ,5 α -THP would be a less convincing model because back-conversion to 5 α -DHP can occur⁹. Furthermore, systemic administration of progesterone best mimics physiological conditions in our animal model of premenstrual syndrome, in which circulating progesterone is converted in the CNS to 3 α ,5 α -THP¹⁰.

Earlier results from Morrow *et al.* indicate that sensitivity to 3 α ,5 α -THP is increased after ethanol withdrawal⁴. We are intrigued by this finding because recent data from our laboratory indicate that withdrawal of 3 α ,5 α -THP also increases ethanol potentiation of GABA-gated current in CA1 hippocampus by $71.2 \pm 5.4\%$ over control levels. This result is consistent with reports of increased alcohol consumption in women during the premenstrual period. Both of these findings clearly distinguish the effects of ethanol from those of benzodiazepines^{1,8}, which exhibit reduced GABA-modulatory

efficacy after 3 α ,5 α -THP withdrawal as a result of upregulation of the GABA-receptor α 4 subunit.

Sheryl S. Smith*, Qi Hua Gong*, Fu-Chun Hsu*, Ronald S. Markowitz*, J. M. H. ffrench-Mullen†, Xinshe Li*

**Department of Neurobiology and Anatomy, Allegheny University of the Health Sciences, EPPI, 3200 Henry Avenue, Philadelphia, Pennsylvania 19129, USA*

†RIN Bioscience, Zeneca Pharmaceuticals, Wilmington, Delaware 19850-5437, USA

- Smith, S. S. *et al.* *Nature* **392**, 926–930 (1998).
- Bjurulf, E., Selstam, G. & Olofsson, J. I. *J. Reprod. Fertil.* **102**, 139–145 (1994).
- Uzunov, D. P., Cooper, T. B., Costa, E. & Guidotti, A. *Proc. Natl. Acad. Sci. USA* **93**, 12599–12604 (1996).
- Devaud, L. L., Purdy, R. H., Finn, D. A. & Morrow, A. L. *J. Pharmacol. Exp. Ther.* **278**, 510–517 (1996).
- Devaud, L. L., Purdy, R. H. & Morrow, A. L. *Alcohol. Clin. Exp. Res.* **19**, 350–355 (1995).
- Devaud, L. L., Fritschy, J. M., Sieghart, W. & Morrow, A. L. *J. Neurochem.* **69**, 126–130 (1997).
- Costa, A. M., Spence, K. T., Smith, S. S. & ffrench-Mullen, J. M. *H. J. Neurophysiol.* **74**, 464–469 (1995).
- Smith, S. S. *et al.* *J. Neurosci.* **18**, 5275–5284 (1998).
- Penning, T. M., Sharp, R. B. & Krieger, N. R. *J. Biol. Chem.* **260**, 15266–15272 (1985).
- Bernardi, F. *et al.* *Eur. J. Endocrinol.* **138**, 316–321 (1998).

Young wallabies get a free ride

Wallabies and kangaroos increase their speed of hopping with little increase in their use of metabolic energy, apparently by means of elastic energy savings in their hindlimb tendons^{1–3}. Here we report that, because of this storage capacity, female tammar wallabies can carry large young in their pouch at very low cost, with no increase in metabolic energy expenditure.

In running mammals, including horses, dogs and rats, the addition of a load results in an increase in metabolic rate that is directly proportional to the load carried⁴. This extra power requirement is presumably linked to increases in mechanical work and to the development of greater force by the muscles. There are a few cases in which this pattern is not seen. Women of the Luo and Kikuyu tribes of Kenya can carry loads equivalent to 20% of their body weight with no measurable increase in metabolic rate, an offset that means they can carry loads of up to 70% of their body weight with less than the expected 70% increase in rate of oxygen consumption^{5,6}. This efficiency is achieved with no increase in mechanical work, as the transfer between kinetic and potential energy components of walking is more complete⁷.

Unlike quadrupeds, in which faster speeds result in linear increases in the mass-specific rate of oxygen consumption, wallabies and kangaroos can increase hopping speed without incurring further increases in metabolic cost^{1,2}. This uncoupling of

metabolic energy requirements and speed is believed to be achieved through the storage and recovery of elastic strain energy in the hindlimb tendons. Potential and kinetic energy lost on landing is stored and subsequently recovered from the tendons during each hop. In wallabies, elastic savings from the tendon represent a significant fraction of the measured metabolic power^{3,8}.

In the tammar wallaby, *Macropus eugenii*, a mother weighing 4–5 kg carries its young in the pouch until its weight approaches 1 kg. Four wallabies were trained to hop on a moving treadmill while carrying a load equivalent to 15% of their body weight. Because of the risk of losing young animals, the load consisted of lead shot placed in a deformable latex glove carried in the pouch. Metabolic rates were assessed from rates of oxygen consumption at two hopping speeds, 3.5 and 4.5 m s⁻¹, in loaded and unloaded

conditions². Direct force recordings were made from the three main ankle plantar–flexor muscle–tendon systems by using stainless-steel buckle transducers³. The time the feet were on the ground and the hopping frequency were obtained indirectly from the recordings of force output.

Adding the load caused no significant change in metabolic rate during hopping, nor did it alter the stride frequency for a given speed, despite the increase in muscle and tendon stress. But adding the load did significantly increase the time for which the foot applied force to the ground during the stride. The duty factor (contact time/stride period) increased from 0.49 to 0.52 at both speeds. The percentage of metabolic power recovered by tendon strain energy increased from 7.2% to 9.2% with the addition of a 15% load at 3.5 m s⁻¹; at 4.5 m s⁻¹ this recovery rose from 9.3% to 12.9%. This increase is less than the 15% increase in mass because the mechanical work saving is given as a percentage of metabolic energy cost. The relationship between these variables is not simple, particularly as duty factor changes. Wallabies carrying a load approximating the mass of a fully developed young have no increase in metabolic energy expenditure, and the increased muscle force required is offset by an increase in stored elastic strain energy. However, the time the foot was on the ground increased in the loaded condition. This might be expected to allow for the recruitment of slower muscle fibres, which could also help lower the energy cost⁹.

Many rodent species and some small birds have adopted a bipedal hopping gait, but kangaroos and wallabies are the only large vertebrates to do so. Given the apparent economies of this locomotory gait, and that the potential for tissue elasticity in muscle–tendon units is strongly mass dependent⁸, this is perhaps surprising. Our findings suggest that the evolution of a bipedal hopping gait has particular adaptive significance in animals that carry their young in a pouch until they reach a stage of advanced body weight. Kangaroos and wallabies may be the only group of animals in which carrying young is energetically 'free'.

R. V. Baudinette*, A. A. Biewener†‡

**Department of Zoology,*

University of Adelaide,

Adelaide, South Australia 5005, Australia

e-mail: rbaudinette@zoology.adelaide.edu.au

†Department of Organismal Biology and Anatomy,

University of Chicago,

Chicago, Illinois 60637, USA

‡Present address: Concord Field Station,

MCZ Department of Organismic and Evolutionary Biology, Harvard University,

Bedford, Massachusetts 01730, USA

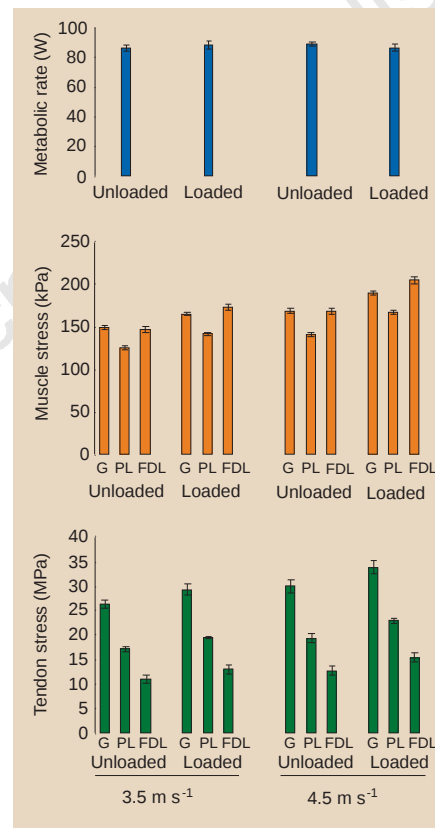


Figure 1 Metabolic rate and muscle and tendon stress in hopping wallabies. Top, metabolic rates, from oxygen consumption measured using a calorimetric equivalent whereby 1 ml O₂ represents 20.1 J, were measured using a flow-through transparent mask held in front of the hopping animal². At the speeds selected, adding a load equivalent to 15% of body weight did not result in significant metabolic changes. Also shown are muscle (middle) and tendon (bottom) stresses, the force per unit cross-sectional area³, in the gastrocnemius (G), plantaris (PL) and flexor digitorum longus (FDL) muscle–tendon units of four wallabies hopping on a moving treadmill. At both hopping speeds, the addition of the load resulted in significant increases in muscle and tendon stress in all three systems ($P < 0.01$).

1. Dawson, T. J. & Taylor, C. R. *Nature* **246**, 313–314 (1973).

2. Baudinette, R. V. *et al.* *Am. J. Physiol.* **262**, 771–778 (1992).

3. Biewener, A. A. & Baudinette, R. V. *J. Exp. Biol.* **198**, 1829–1841 (1995).

- Taylor, C. R., Heglund, N. C., McMahon, T. A. & Looney, T. R. *J. Exp. Biol.* **86**, 9–18 (1980).
- Maloij, G. M. O., Heglund, N. C., Prager, L. M., Cavagna, G. A. & Taylor, C. R. *Nature* **319**, 668–669 (1986).
- Charteris, J., Scott, P. A. & Nottrodt, J. W. *Ergonomics* **32**, 1539–1550 (1989).
- Heglund, N. C., Willems, P. A., Penta, M. & Cavagna, G. A. *Nature* **375**, 52–54 (1995).
- Bennett, M. B. & Taylor, G. C. *Nature* **378**, 56–59 (1995).
- Kram, R. & Taylor, C. R. *Nature* **346**, 265–267 (1990).

Stabilizing gaze in flying blowflies

When the gaze is moved quickly, vision becomes blurred and the way in which three-dimensional structures are seen can be affected¹. How can these effects be minimized? During flight, blowflies (*Calliphora vicina*) turn both the head and the thorax very quickly, producing gaze shifts which affect vision. Here we show that blowflies reduce the effects of thorax movements on vision by moving the head later, and more quickly, than the thorax. This reduces the time in which gaze is shifting relative to the surroundings, and maximizes the time available for analysing the surroundings.

Gaze shifts present a challenge to vision for at least two reasons. First, it takes time for optic signals to be integrated by the photoreceptors in the eye. If the gaze is shifted, this integration time will result in a blurred image being produced, particularly during fast gaze shifts. Second, the resulting rotational flow of optic signals will interfere with the pattern of optic flow that normally reveals the three-dimensional structure of the surroundings during linear motion¹.

Blowflies can move their two compound eyes by moving the head^{2,3}. The head, which is connected to the thorax by a neck, has appreciable freedom of movement under muscular control^{4,5}. We measured the position and orientation of thorax and head during (almost) free flight by using a modified search coil technique⁶. The thorax usually changes course through a series of short, fast turns⁷, with gradual course

changes occurring less frequently. The head usually turns in synchrony with the thorax but at a higher angular speed (Fig. 1a), with a pattern similar to that of fast eye movements in vertebrates⁸.

The average angles of yaws and rolls (Fig. 1b) during head–thorax turns of 10–20 degrees to the left (the most common yaw angle; turns of other angles follow similar time courses) are shown in Fig. 1c,d. A turn starts with a rotation of the thorax. The thorax yaw (Fig. 1c) is not immediately accompanied by a head yaw relative to the surroundings, as it is first compensated by a rotation of the head in the opposite direction relative to the thorax. After about 10 milliseconds, the head starts to rotate in the same direction as the thorax. It reaches its highest yaw velocity relative to the thorax at about the same time as the thorax reaches its highest yaw velocity relative to the surroundings. When the head approaches its final orientation relative to the surroundings, it decelerates, and finally rotates in the opposite direction to the thorax. The result is a stabilized yaw of the head roughly 10 milliseconds before the thorax rotation is finished. As a result of the shortening of the head turn compared with the thorax turn, the period of blurred vision during course changes is reduced by roughly one-third.

Like aeroplanes, flies usually make banked turns⁷, in which large changes in yaw are accompanied by large rolls of the thorax. The head can partly correct the visual consequences of this by performing a counter roll⁹ (Fig. 1d). The residual roll movement of the head relative to the surroundings is small (typically less than 5 degrees). Rotations in the pitch direction (not shown) are also usually small.

We can deduce the visual consequences of coupled thorax and head turns from Fig. 2. We divided the total flight time into episodes during turns and episodes between turns. During turns, the yaw velocity of the head is higher than that of the thorax (Fig. 2a). As the photoreceptor integration time for the light level during the measurements was roughly 10 milliseconds,

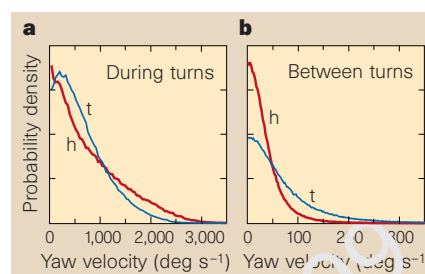


Figure 2 Gaze stabilization between turns. **a**, Probability densities of yaw velocities during turns, showing that yaw velocities of the head (h) reach higher values than those of the thorax (t). Total flight time was 703 seconds by four flies, during which time 6,697 turns were made. **b**, Between turns, yaw velocities of the head are significantly lower than those of the thorax, and both are much lower than during turns. More detailed figure legends are available as Supplementary information.

and the angular resolution of the photoreceptors is 1–2 degrees, angular velocities higher than 100–200 degrees per second will cause significant blur in the fly's visual system. The high angular velocity of the head during turns therefore makes it impossible to see details. Between turns, however, the head is more stable than the thorax (Fig. 2b), mostly within a velocity range that gives little blur attributable to rotation.

The triphasic movement of the head in the yaw direction (Fig. 1c) probably acts to maximize the periods of stable gaze by minimizing the duration of the gaze shift. The extra angular velocity produced by the neck muscles is effective because it peaks at about the same time as the peak of the thorax angular velocity. The effective duration of the most common gaze shifts is not much longer than the integration time of the photoreceptor. Because at least one integration period is lost by making a gaze shift anyway, the roughly 1.5 integration periods lost during a typical blowfly turn seem to be a reasonable compromise between minimal visual impediment and very fast movement.

C. Schilstra, J. H. van Hateren

Department of Neurobiophysics,
University of Groningen,

Nijenborgh 4, 9747 AG Groningen, The Netherlands
e-mail: hateren@bcn.rug.nl

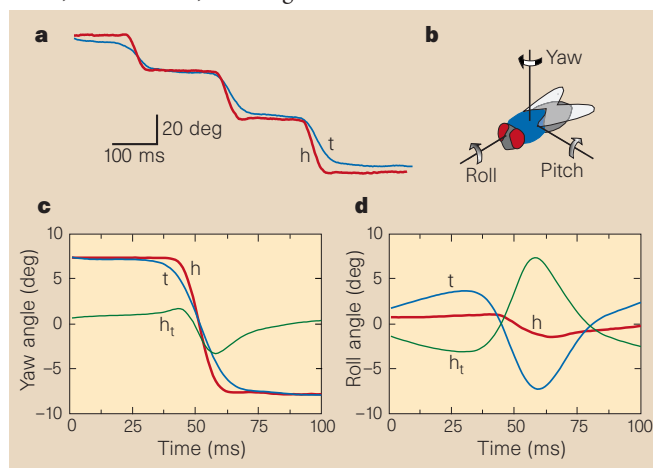


Figure 1 Turns in blowflies.

a, Turns in the yaw direction of the thorax (t) are joined by faster turns of the head (h); angles relative to surroundings. **b**, Rotations have an ordered sequence of yaw, pitch and roll (Fick coordinate system¹⁰); signs comply with aeronautical conventions. **c**, **d**, Yaw or roll of thorax and head in turns of 10–20 degrees (average of 713 turns of 4 flies); h_t shows yaw or roll of head relative to thorax.

- Koenderink, J. J. *Vision Res.* **26**, 161–180 (1986).
- Land, M. F. *Nature* **234**, 299–300 (1973).
- Geiger, G. & Poggio, T. *Biol. Cybernet.* **25**, 177–180 (1977).
- Hengstenberg, R. in *The Head-Neck Sensory Motor System* (eds Berthoz, A., Graf, W. & Vidal, P. P.) 49–55 (Oxford Univ. Press, 1992).
- Strausfeld, N. J., Seyan, H. S. & Milde, J. J. *J. Comp. Physiol. A* **160**, 205–224 (1987).
- Schilstra, C. & van Hateren, J. H. *J. Neurosci. Methods* (in the press).
- Wagner, H. *Phil. Trans. R. Soc. Lond. B* **312**, 527–551 (1986).
- Carpenter, R. H. S. *Movements of the Eyes* (Pion, London, 1988).
- Hengstenberg, R. *J. Comp. Physiol. A* **163**, 151–165 (1988).
- Haslwanter, T. *Vision Res.* **35**, 1727–1739 (1995).

Supplementary information is available on Nature's World-Wide Web site (www.nature.com) or as paper copy from the London editorial office of Nature.

The bird-watcher's guides to paradise

A Guide to the Birds of Wallacea

by Brian J. Coates, K. David Bishop and D. Gardner

Dove Publications, Box 59, Alderley, Queensland 4051, Australia: 1997. 535pp. A\$65, £38.50, US\$50

The Birds of Paradise

by Clifford B. Frith and Bruce M. Beehler
Oxford University Press: 1998. 613pp. \$85

Jared Diamond

The great nineteenth-century British naturalist Alfred Russel Wallace is now widely credited with co-discovering natural selection and with having single-handedly founded the science of biogeography. He is also credited with having explored much of Indonesia biologically, in the face of obstacles that remain daunting today. That Wallace emerged from Indonesia alive in 1862 was nothing short of miraculous. Only in recent decades have field biologists succeeded in adding significantly to our knowledge of some of his discoveries. Two recent books by modern biological explorers provide massive compilations of some of this new knowledge.

One of these two books — that by Coates, Bishop and Gardner — could be termed a field guide to the birds of Central Indonesia. But that description would miss the book's broader significance. Central Indonesia between the Asian and Australian continental shelves is the most drastic biogeographic transition zone in the world: a region where tigers and pheasants yield to kangaroos and birds of paradise. It is no accident that this is the region where the search for explanations brought Wallace his insights into biogeography. In his honour, biologists now refer to the region as Wallacea.



Birds of paradise, clockwise from top left: King of Saxony, Magnificent, King, Wilson's and Superb.

Guidebooks to birds were previously available for the regions flanking Wallacea to the west (the Greater Sunda Islands) and to the east (Australia and New Guinea), but not for Wallacea itself. In the absence of a Wallacean field guide, ornithologists contemplating field studies there had to prepare themselves by spending weeks in museums with large bird collections and virtually assembling their own guide. Not surprisingly, few ornithologists have chosen to work in Wallacea. The new field guide at last makes ornithological field work there practical.

With 697 bird species, of which 249 are endemic to Wallacea, the region is very rich. It consists of thousands of islands falling into three groups: the Lesser Sunda chain from Lombok to Tanimbar, along which the transition from an Asian to an Australian fauna is most diagrammatic and linear; the large, weirdly shaped island of Sulawesi (alias Celebes), with four long thin peninsulas resulting from crustal plates crashing together from several directions; and the Moluccas off the western end of New Guinea. Altitudes varying by more than 3,000 metres, and habitats ranging from savannah through rainforest to alpine grassland, make several of the islands as ecologically diverse as miniature continents.

For most of Wallacea's endemic bird populations, the book provides the first published descriptions of songs and behaviour, derived largely from the authors' own field work. Hence the book has the form of a fat field guide (535 pages and 64 colour plates) but the significance of a monumental research treatise.

Two examples must suffice to illustrate the hundreds of potential PhD thesis topics now awaiting young ornithologists in Wallacea. Plate 47 depicts Wallace's greatest single discovery, the Standardwing Bird of Paradise of the Northern Moluccas, rediscovered in 1983 by the book's second author K. D. Bishop. The Standardwing turns out to be a lek breeder whose males, on the approach of a female, go into a frenzy, leap vertically up to 11 metres into the air, and descend parachute-like with their four long white wing standards flared. What information do females extract from this display that could help them choose the best mate? Plate 60 depicts the large honey-eaters known as friarbirds, whose geographic variation of plumage among islands is duplicated in stunning detail by the smaller orioles depicted on Plate 46. Was Wallace right in inter-



Turning tail: Blue Birds of Paradise; males (top) showing dual-coloured tail plumage from front and back.

preting this as a case of visual mimicry of friarbirds by orioles, and, if so, what advantage do the orioles thereby gain?

Frith and Beehler's book, at 613 pages equally monumental, is an exhaustive summary of literally everything known about all 42 birds of paradise (BOPs), a family confined to New Guinea and Australia except for two species in Wallacea. All those BOPs are illustrated in colour plates by the famous painter William Cooper.

Male BOPs are the world's most bizarrely plumaged birds. For instance, from behind each eye of the thrush-sized King of Saxony Bird of Paradise sprouts a wire-like feather shaft 50 cm long, adorned on one side with up to 50 square blue 'flags' resembling pieces of plastic.

When the first specimen reached Europe in 1894, the museum curator R. Bowdler Sharpe wrote that any fool could recognize the alleged bird as a glued-together artifact. In fact, the supposedly glued-on bits of plastic, and the equally bizarrely modified plumes of other BOP species, function in courtship displays, raising in extreme form

the questions that pervade the field of mating biology in general.

BOPs prove ideal subjects of investigation for many other biological questions. Through the lens of BOP biology, Frith and Beehler's book summarizes such frontiers of biology as animal/plant mutualism, body toxins, diet, evolutionary radiation, heterochrony, hybridization, mating systems, mimicry, mixed species flocks, nesting, phylogeny, sexual dimorphism, skull modification and speciation.

Again, a set of examples from the authors' information summary and provocative unanswered questions must suffice. Most BOPs are predominantly frugivores as adults, a characteristic usually considered as predisposing them to evolve their famous court-displaying polygyny. But what permits the Buff-tailed Sicklebill as an exception to combine court-displaying polygyny with nearly strict insectivory? Nestlings of most BOP species are instead fed arthropods, presumably to provide the protein and fat that they need for growth. But how do nestling manucodes and at least three other BOPs manage to subsist on fruit? Their iron-poor fruit diet may explain why captive BOPs fed iron-rich zoo diets tend to die of iron poisoning. Could BOPs thus serve as a model for genetically-based iron poisoning — haemochromatosis — in humans? As specialized frugivores, BOPs constitute the most important group of vertebrate seed dispersers in New Guinea, being especially adept at using their feet and bills together to extract fruits from a protective capsule. Did this adaptation evolve because of New Guinea's lack of squirrels and monkeys, which pre-empt such fruit manipulation elsewhere in the world? Why is one of the two BOP subfamilies unskilled at fruit manipulation? BOP/fruit relations are among the most species-specific fruit-foraging systems known, but does this represent a mutualism or just a one-sided specialization on the plants' part?

I must also comment on Frith and Beehler's organizational skill, from which any would-be author could learn. Masses of details are rendered digestible by presenting eight introductory chapters on general biological topics, then 42 more specific bird-by-bird accounts, and finally seven technical appendices. Information about unresolved questions escapes becoming a mass of detail by being organized around unifying hypotheses, but competing hypotheses are given fair statements and evaluations. Questions for future research are highlighted throughout, making this book another treasure trove for students in search of PhD thesis topics.

Who should buy these books? Neither has a low price-tag, but both are bargains when measured by the huge amounts of information they contain that is unavailable else-

where. Any birder or biologist contemplating a visit to Indonesia needs Coates, Bishop and Gardner; and any birder or biologist not contemplating a visit to Indonesia may rethink plans after seeing the book. Biologists interested in phenomena from animal/plant mutualism to speciation — especially those in search of a new research project — will profit from Frith and Beehler. So will any scientist seeking inspiration on how to organize a book on picosecond chemistry, molecular genetics, or other complex fields of science remote from avian biology. □

Jared Diamond is in the Physiology Department, University of California Medical School, Los Angeles, CA 90095-1751, USA.

Are talkers the only thinkers?

If a Lion Could Talk: Animal Intelligence and the Evolution of Consciousness

by Stephen Budiansky
Free Press: 1998, 219pp, \$25

Karen McComb and Stuart Semple

"One chimp was alone in the feeding area and was going to be fed bananas. A metal box was opened electrically from a distance. Just at the moment when the box was opened, another chimp approached at the border of the clearing. The first chimp quickly closed the metal box and walked away several metres, sat down and looked around as if nothing had happened. The second chimp left the feeding area again, but as soon as he was out of sight, he hid behind a tree and peered at the individual in the feeding area. As soon as that individual approached and opened the metal box again, the hiding indi-

vidual approached, replaced the other and ate the bananas."

This intriguing observation of the Gombe chimpanzees by Frans Plooij (Whiten, A., Byrne, R. W. *Primate Report* 27, 99, 1990) can be interpreted in two very different ways. Perhaps the chimpanzee that hid himself was capable of attributing a mental state — the intention to deceive — to his social companion and made use of this information in order to outwit him. Alternatively, as an extremely astute observer, he could simply have learnt by trial and error that this intricate pattern of behaviour was likely to secure him a reward.

The problem of distinguishing between these two sorts of explanation is by no means trivial and it applies to almost every study in which behaviour is used to assess the nature of the animal mind.

Budiansky is therefore right to explore in depth the feats of "unthinking intelligence" that associative learning can potentially produce, and to examine how this process may underlie behaviours which initially suggest sophisticated reasoning on the part of the animals involved. To his credit, he does this lucidly and with a wealth of examples drawn from throughout the animal kingdom. He also highlights the important fact that cross-species comparisons of mental abilities are often biased towards species that have inherent predispositions to learn how to perform the laboratory tests involved. This point is particularly salient given the relative ease with which certain primates adapt to interacting with humans in the laboratory and to performing tests that involve elements of manual dexterity.

In light of these problems, how does Budiansky address the central issue of the book — the question of whether animals are capable of 'conscious' thought? His attempts



Buster Keaton and comrades in *The Doughboys*: is our understanding of animal intelligence distorted by anthropomorphism?

to do this are seriously hampered by the fact that he never provides us with a clear definition of what he considers 'consciousness' to be. One infers as the book unfolds that consciousness, for him, involves second order intentionality (in Daniel Dennett's sense) — the ability to have thoughts about thoughts. First order intentionality, the ability to have thoughts in the first place, is not sufficient.

But Budiansky goes further than this. He appears slavishly caught in the stance that the only mechanism through which conscious reasoning can occur is human language, which he believes "is so intimately tied to consciousness that the two seem inseparable". This obviously excludes the possibility that animals could have non-verbal thoughts about thoughts, a dangerous assumption given that studies on human infants suggest that elements of second order intentionality can in fact occur in the absence of language.

The conclusion from Budiansky's viewpoint can only be that animals, lacking the linguistic abilities of humans, aren't conscious.

What, then, are we to make of evidence that a variety of primates use representational alarm calls to denote specific classes of predator; that species as wide-ranging as African grey parrots, bottlenosed dolphins and pygmy chimpanzees readily adapt to using acoustic or visual labels to represent external objects and events; that rhesus monkeys can represent the ordinal position of objects in arbitrary lists; or that vervet monkeys have representations for certain sorts of social relationships?

At one point Budiansky admits that "the question of what a mental representation is lies at the heart of the mystery of animal thinking and consciousness". But direct discussion of this issue in the light of findings that animals do appear to use mental representations is sadly lacking. This can only partly be attributed to Budiansky's reluctance to accept the results of these studies. Although he dismisses evidence for representational alarm calls (even though their existence is widely accepted by experts in the field) and is singularly unimpressed by the achievements of lexigram-trained apes, he does consider that vervet monkeys "have no trouble spontaneously mastering mental representations of social relationships" and that "monkeys really have an ability to store items in sequence and mentally run through and draw conclusions from such a sequential list". It is frustrating that Budiansky simply seems unwilling to examine the full implications of these rudimentary abilities to mentally code information on things external to self.

Given the nature of Budiansky's approach it is perhaps unsurprising that he should advocate abandoning studies of animal consciousness, in favour of investigating things that animals "are good at". His judgements about what animals should be good at —

"things they have evolved to do to survive in their particular ecological niche" — seem to be extremely limited. For example, he appears willing to include under this umbrella abilities for hiding and recovering food but not a capacity for counting, even though social animals such as lions can and do benefit from accurately assessing the number of opponents that they face in territorial disputes. He also fails to appreciate that the complex social environment in which many animals live would be likely to exert strong selective pressures for the evolution of an ability to attribute mental states. In particular, the far greater predictive power afforded by reading the minds of others rather than simply their behaviour would provide a much more effective means for individuals to outmanoeuvre social competitors.

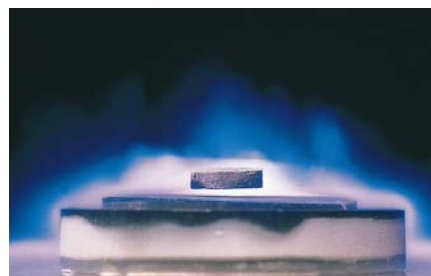
It is Budiansky's strongly stated belief that our motivation for studying consciousness is biased by an anthropocentric view of the world. Perhaps this is why he feels that one should not seek human-like abilities, such as the understanding of mental states, in animals. However, this conclusion is unjustified. The question of whether any species other than our own can appreciate that it is possible to manipulate minds as well as behaviour is no mere anthropocentric indulgence — instead it is one of the most important and challenging issues in social evolution.

In light of his convictions about the inextricable link between consciousness and language, it is probably no coincidence that Budiansky's book is called *If a Lion Could Talk*. Indeed, he suggests that if a lion *could* talk "his mind would no longer be a lion's mind". The fact that an animal cannot talk should not, however, be taken to indicate that it lacks consciousness — it is absolutely crucial not to prejudge the issue. Unfortunately, Budiansky appears to have done just this.

Even if consciousness is equated to second-order intentionality as Budiansky proposes, the problems of finding appropriate methods to discern whether animals can attribute a mental state to others are not insurmountable. Although animals lack our language, their understanding of the intentions and knowledge of others could still be examined if more imaginative experimental paradigms were employed. The techniques used to monitor the development of mental state attribution in human infants are potentially very promising in this respect.

It is disappointing that, far from suggesting new directions the study of animal consciousness could take, Budiansky actively discourages further research in this fascinating area. □

Karen McComb is in the Laboratory of Experimental Psychology, School of Biological Sciences, University of Sussex, Brighton BN1 9QG, UK. Stuart Semple is at the Institute of Zoology, Regent's Park, London NW1 4RY, UK.



High hopes: a magnet floating above nitrogen-cooled superconducting ceramic.

Bound for the future

Handbook of Applied Superconductivity

edited by Bernd Seeber
Institute of Physics: 1998. 2 vols, 1,912pp.
£350, \$545

David Larbalestier

Late in 1986, a tentative report of the discovery of superconductivity at temperatures above 30 K in a strange multi-component copper oxide was independently confirmed by several groups. Within three months, multiple laboratories, even a few high schools, had "touchable" demonstrations of the levitational marvels of persistent current flow, due to these copper oxide, high-temperature superconductors (HTS) that worked in liquid nitrogen. An imminent superconducting age was widely predicted, mostly by those who knew little of the existing low-temperature superconducting (LTS) industry. Now, 11 years later, comes a 1,900-page *Handbook of Applied Superconductivity*. Its coverage provides a fascinating insight into the present state of superconducting technology and, at least by implication, poses several key questions about the eventual applications of HTS.

The handbook, largely written by experienced Europeans, shows that the technology of superconductivity rests on a sophisticated base. However, much less than 20 per cent of the book is concerned with high-temperature superconductivity — the copper oxide superconductors that the public thinks the applications of superconductivity are all about. The book presents a detailed view of how to design and construct superconducting magnet systems (only a small part of the book, about 150 pages, examines electronic applications), paying greatest attention to the conductor (about 425 pages), magnets (425 pages) and refrigeration (290 pages).

For those wanting the whole story, the handbook is the finest present sourcebook because of its detail, its breadth of authorship and its detailed bibliography. I was particularly impressed by the detailed coverage of all aspects of wire and cable design, the vital centrepiece of all superconducting devices.

What of those looking for clues as to when

HTS applications will come to market? Had this been written by American or Japanese authors, there might have been a greater discussion of HTS applications (though these do figure strongly in the 250-page section on power applications). But it is as a snapshot of what today's LTS-based industry makes — magnetic resonance imaging magnets, particle accelerators and various sorts of laboratory magnets — that the handbook excels.

The handbook can be profitably studied to understand why HTS will not easily displace LTS from present applications. By far the biggest present market is for magnetic resonance imaging magnets, which all use LTS conductors, partly because LTS Nb–Ti conductors are so good and cheap, and partly because their sophisticated cryostat design demands only an annual helium filling. HTS current leads, the first killer application of HTS, actually make the task of replacing all LTS conductors in present applications unnecessary, for well-designed HTS current leads can reduce the heat leak to liquid helium by about 90 per cent, permitting helium-free ('dry') electromagnets wound with LTS wire. Widespread HTS applications await better, cheaper conductors for power cables, transformers, motors and other electric power applications.

Recent HTS conductor progress has been impressive, but a handbook like this shows that LTS conductors are still far from being counted out in the continued drive for more superconducting applications. □

David Larbalestier is in the Applied Superconductivity Center, University of Wisconsin, 1500 Engineering Drive, Madison, Wisconsin 53706, USA.

When the boat comes in

Marine Biodiversity: Patterns and Processes

edited by R. F. G. Ormund, J. D. Gage and M. V. Angel
Cambridge University Press: 1997. 442pp.
£50, \$74.95

Linda Medlin

Biodiversity is the environmental concern of the 1990s and will remain a concern into the next millennium, with ecosystem destruction and species extinction now recognized on a global, epic scale. For some time, marine biologists and ecologists have stressed that far less is known about the biodiversity of marine environments than that of terrestrial ecosystems. Indeed, the processes of speciation and dispersal in the two ecosystems are likely to be vastly different. *Marine Biodiversity* helps to redress the balance by explaining major patterns of diversity in the marine environment.

Chapter subjects range from diversity in



Fragile beauty: *Anthias pseudanthias* tuka coral in the waters off Bunaken Isle, Indonesia.

the upper waters of the open sea (the pelagic) to that in shallow seas, and from the bottom of deep seas (the benthic) to coral reefs. The writers consider how historical, biogeographic and oceanographic factors have structured biodiversity at the α (local habitats), β (between habitats) and γ (regional) levels, and how this biodiversity is maintained and should be managed. Of special interest is the evidence for nutrient availability controlling tropical Indo-Pacific benthic community diversity; for large-scale current disturbances affecting deep-sea diversity; and for the control of diversity in estuaries by coastal geomorphology. Notably absent, however, are any chapters devoted to phytoplankton or macro-algal diversity, a common oversight in studies on biodiversity.

Commonly held ideas about marine biodiversity, such as the existence of large randomly breeding populations, are gradually loosening their hold. Molecular genetic studies of marine populations continue to reveal small genetically effective population sizes, making more careful management of marine populations essential. Gradients in marine diversity, especially latitudinal ones, are repeatedly called into question for certain geographical areas, such as the southern hemisphere, and for certain ecosystems, such as the shallow to deep sea benthos. Because these themes are discussed in several chapters, there is some overlap in the material, and more cross-referencing could have been made. Historical (geological) legacies play an important role in shaping biodiversity and should be used more to counter current views of the predominance of latitudinal gradients in the marine environment.

No book on biodiversity would be complete without a discussion of molecular

genetic variation in populations and how this can be interpreted in terms of local adaptation and isolation, which could lead to speciation. The lack of resolution in population boundaries suggested by allozyme data is continually contrasted to that recovered with sequence data, so problems associated with identifying cryptic species still persist. With few hints of morphological variation in some species, small amounts of genetic variation tend to be overlooked. These differences are now being appreciated, thanks to the rapid increase in the number of sibling species recognized in the marine environment.

The Baltic is an example of a low-diversity area, and studies in this water body provide the clearest evidence for how ecosystem function can be threatened if a major functional group is eliminated from the community. Ecological functioning of biodiversity is explored in the contribution by G. C. Ray *et al.*, who propose a landscape pattern approach from the ecosystem down to the species level. They predict that the decline or removal of a species will influence both total biodiversity and environmental conditions because feedback mechanisms are removed. Clearly, this would have far-reaching effects for ecosystem management. Although, as pointed out by J. S. Gray, the 'hotspots' of greatest diversity have been targeted for monitoring by international agencies, these are not the areas that are subject to man's greatest influence.

Mariculture also has an effect on biodiversity, but is not often highlighted. Its impact on biodiversity can be both beneficial and harmful, so careful planning is vital.

The overriding themes emerging from this book are that we still know very little about the factors controlling marine biodiversity in certain ecosystems, and far less about the mechanisms controlling genetic diversity. However, because of the major influence that the marine ecosystem has on our global climate, and because of the heavy reliance man has on marine resources, the impacts of disturbances on marine biodiversity should not be underestimated.

Most authors call for more quantitative data to assess biodiversity at all levels in the marine ecosystem before effective management can be implemented, but we mustn't wait too long. Species are disappearing at an alarming rate. The book even suggests that perhaps we need some catastrophe just large enough to persuade governments and peoples to look at ocean management in a more radical fashion, rather than simply for the protection of certain species, the disposal of particular wastes, or the study of unusual phenomena. Effective ocean management to sustain and protect marine biodiversity is a global issue that is here to stay. □

Linda Medlin is in the Department of Biological Oceanography, Alfred Wegener Institute, Am Handelshafen 12, D-27570 Bremerhaven, Germany.

A vertebrate model of extreme physiological regulation

Stephen M. Secor & Jared Diamond

Department of Physiology, University of California Medical School, Los Angeles, California 90095-1751, USA

Investigation of vertebrate regulatory biology is restricted by the modest response amplitudes in mammalian model species that derive from a lifestyle of frequent small meals. By contrast, ambush-hunting snakes eat huge meals after long intervals. In juvenile pythons during feeding, there are large and rapid increases in metabolism and secretion, in the activation of enzymes and transporter proteins, and in tissue growth. These responses enable an economic hypothesis concerning the evolution of regulation to be tested. Combined with other experimental advantages, these features recommend juvenile pythons as the equivalent of a squid axon in vertebrate regulatory biology.

The history of biology illustrates the importance of selecting exceptionally suitable species as models: examples include the contributions made by squid axon, pigeon breast muscle, *Necturus* kidney and *Drosophila* to our understanding of excitable membranes, oxidative metabolism, kidney function and population genetics, respectively. These models distinguished themselves by exaggerated structures or responses, or by experimental convenience with respect to the phenomenon under study. Once biologists had unravelled the phenomenon in the model species, they could devise experiments for doing so in species (such as humans and other mammals) presenting greater experimental difficulties but with more practical importance.

A field now in need of such a model species is vertebrate regulatory biology. Many processes are regulated on various time-scales by food intake: rapid responses include release of gastrointestinal hormones, switching-on of gastrointestinal secretions, upregulation of nutrient transporters and hydrolases in the gut, and a rise in metabolic rate to accompany digestion; slow responses to chronically increased food intake (for example, during lactation, accompanying an athletic lifestyle, or for heat production in a cold environment) include growth of the heart, kidney, liver and intestine, and adaptations of pulmonary and cardiac function.

Humans and the usual mammalian model species (rats, mice and rabbits) are adapted to consuming small meals (equivalent to only a few per cent of body mass) frequently (many times daily). Hence the gut usually contains food, and fluctuations in loads upon metabolic processes are modest. As a result, regulatory responses have evolved to encompass only modest factorial spans (Table 1), making them difficult to study experimentally, despite their physiological and clinical importance.

A new vertebrate model that exhibits much larger regulatory responses would be valuable for two reasons. One reason is to advance our understanding of regulatory phenomena, such as organ growth and atrophy or signal pathways for hormone release. The other reason is to understand the evolution of regulation itself. Although many biological parameters are regulated reversibly, others are fixed and unresponsive to changes in conditions. For instance, the masses of the mammalian kidney and intestine vary with food intake, but those of the pancreas or brain do not; likewise, the activity of intestinal sucrase, but not of erythrocyte enzymes, varies with food intake. Could these contrasting outcomes depend on the relative costs of maintaining a component 'ready to go' compared with synthesizing it only when needed, and also depend on the temporal variation in component operation typically required over an animal's adult life? (By analogy, someone driving a car in normal traffic finds it cheapest to keep the car's engine running while stopping briefly at traffic lights, but turns off the

engine, thus saving fuel consumption, and restarts it after stopping at a railroad while waiting for a long train to pass.) Testing this evolutionary hypothesis requires a system in which the relative costs of maintaining compared with periodically synthesizing certain biological machinery can be separately identified and measured.

We sought such a model among animal species that, unlike rats and humans, are adapted to consuming large meals at long, erratic intervals. Examples of such species include lions, wolves and some deep-sea fishes. As experimentally more tractable candidates, we tested eighteen species of frogs, lizards, turtles and snakes that were reputed to consume large meals. We measured maximum voluntary meal size and post-feeding increases in metabolic rate, enzyme and transporter activities, and organ masses.

The most promising candidates proved to be snake species that obtain their prey by waiting in ambush. Field studies of such snakes show that the mass of a prey animal, swallowed whole without chewing, averages one-quarter of the snake's body mass but ranges up to 1.6 times the snake's mass^{1,2}. (That is analogous to a person weighing 62 kg swallowing a 100 kg meal in one gulp.) Typical feeding intervals for such snakes in the wild are one or two months, but may exceed one year^{1,3}. Our eventual choice of species was the Burmese python, *Python molurus*. Adults are among the largest snakes, reaching 6.5 m long and 100 kg, and they consume large mammals (including humans)³. However, juveniles weigh only 0.1–1.0 kg, consume rats and mice, are popular pets, are available from commercial breeders, and as experimental animals possess other virtues that we shall describe.

Table 1 Comparison of regulatory spans in pythons and mammals

	Factorial increase	
	Pythons	Mammals
Post-feeding response		
Kidney mass	2.1	1.1 (m)
Intestinal mucosal mass	2.2	1.6 (m)
Plasma glucose	2.3	1.2 (h)
Plasma free fatty acids	2.5	1.5 (h)
Intestinal maltase activity	3.0	1.3 (r)
Intestinal peptidase activity	5.0	1.8 (r)
Intestinal microvillus length	6.0	1.6 (ha)
Intestinal amino acid transport rates	10	2.0 (m)
Intestinal glucose transport rates	41	1.7 (m)
Plasma insulin	41	5.0 (h)
Metabolic rate	44	1.5 (h)
Plasma cholecystokinin	52	6.5 (h)
Plasma triglycerides	160	1.7 (h)

Numbers are post-feeding regulatory spans of various quantities: that is, the factorial increase from fasting levels to peak levels after feeding. For instance, plasma triglyceride levels rise by a factor of 160 in pythons but by only 1.7 times in humans. Note that all regulatory spans are much greater in pythons than in well studied mammal species (m, mice; ha, hamsters; r, rat; h, humans). Sources: for pythons, refs 4 and 8 and personal observations; for mammals, published references available upon request.

Regulatory responses to feeding

Upon swallowing its prey, a python coils up and remains nearly motionless (except for breathing deeply) throughout the 5–11 days required to complete digestion (larger meals requiring more days). This apparent inactivity conceals vigorous internal metabolic activity. Table 1 summarizes factorial magnitudes of some underlying regulatory responses and compares them with the much smaller responses of humans or rodents. These responses all revert to fasting levels by the time of defecation.

A python's metabolic rate, as measured by rates of oxygen consumption ($\dot{V}_{O_2}$), rises within 3 h of feeding, peaks at 1–2 days, and declines to fasting levels at 4–16 days (Fig. 1). Peaks in $\dot{V}_{O_2}$ are up to 44-times fasting $\dot{V}_{O_2}$ in snakes digesting meals equal to the snake's own body mass⁴ (Fig. 1). For comparison, the highest factorial post-feeding $\dot{V}_{O_2}$ rise reported for any digesting mammal is 2.0, for dogs⁵. The highest factorial $\dot{V}_{O_2}$ peak reported for a mammal under any conditions is 45 for a galloping racehorse, which is virtually the same as for a digesting python. However, horses can sustain a gallop for only a few minutes, whereas digesting pythons maintain elevated $\dot{V}_{O_2}$ for up to 2 weeks. Pythons' high factorial peak $\dot{V}_{O_2}$ is due not only to their high absolute $\dot{V}_{O_2}$ during digestion, but also to their low standard (fasting) metabolic rate, which is about 13 times lower than that of a similarly sized mammal at the same body temperature⁶.

Pythons' increased $\dot{V}_{O_2}$ during digestion arises from metabolic costs of regulatory processes such as switching-on gastrointestinal secretions, upregulating enzymes and transporters, and stimulating rapid growth of organs. The secretion of hydrochloric acid by the stomach causes gastric pH to plummet from 7 to 1 within one day, and to remain low for one week (compared with only a few hours in humans). In the intestinal brush border, the activities of the enzyme amino-oligopeptidase and of the glucose and amino-acid transporters rise by a factor of 5–40 within 1–3 days^{7,8}. The small intestine doubles in wet and dry mass within 1 day, largely as a result of a sixfold increase in microvillus length and a doubling of mucosal enterocyte volume (Fig. 2). Within 1–3 days, there are also 50–100% increases in masses of the stomach, liver, pancreas, heart, lungs and kidneys (Fig. 3).

While growth of the small intestine, stomach, liver and pancreas is obviously related to their role in digestion, it is initially surprising that the heart, lungs and kidneys grow as well. However, these organs too experience increased work loads during digestion. Increased O_2 consumption (Fig. 1) and CO_2 exhalation require heart rate, blood flow and ventilation to increase by 3–5 times (S.M.S., J. Hicks and A. Bennett, unpublished observations), whereas increased production of metabolic waste requires increased

renal excretion. All of these organs atrophy back to fasting levels after defecation.

Organ growth and enzyme and transporter synthesis require biosynthetic energy. Yet they are underway or have peaked at one day after feeding, when the swallowed rat is still largely intact within the snake's stomach and when intestinal digestion and absorption have scarcely begun. Evidently, energy and substrates for growth and synthesis must initially be mobilized from the snake's stored energy reserves, not from the rat's body. This mobilization is reflected within a day in a 160-fold rise of plasma triglycerides, possibly originating from the large, paired fat bodies within the python's body cavity and causing the plasma at day 1 to change colour from clear to milky-white. That is, digesting pythons operate on the principle of many North American self-service petrol stations: pay before pumping.

The evolution of regulation

The high $\dot{V}_{O_2}$ of digesting pythons reflects their high costs of digestion. In part, these high costs are due to pythons' exceptionally large meals: their total cost of digestion (also known as specific dynamic action (SDA)), which is calculated as $\dot{V}_{O_2}$ beyond the standard metabolic rate and integrated over the period of digestion, increases linearly with meal size⁴. However, their relative cost of digestion is also high for any meal size: it represents on average 32% of the ingested meal's energy equivalent⁴. This percentage, termed the SDA coefficient, is only 9% for humans and falls between 4% and 17% for most vertebrate species^{4,9}. That is, digesting pythons burn up a much higher fraction of their meal's energy-equivalent content than do humans.

This high relative cost of digestion is surely related to the biosynthetic start-up costs that digesting pythons incur in synthesizing organs and proteins that have atrophied or been repressed during fasting. Humans and other frequent feeders are spared the start-up costs because they maintain those organs and proteins constantly. Conversely, pythons spare themselves much of the high maintenance cost of those organs and proteins incurred in frequent feeders, by letting them regress after digestion is complete. For instance, the small intestine, heart, liver and kidneys are among the organs with the highest metabolic rates, so that they contribute to only one-eighth of a rat's body mass, but one-quarter to its basal metabolism¹⁰.

Did pythons evolve these extreme feeding-related regulatory responses because of their very infrequent meals? That is, are energy budgets that include the occasional high start-up costs associated with gut upregulation lower than energy budgets that include high everyday maintenance costs such as would be incurred

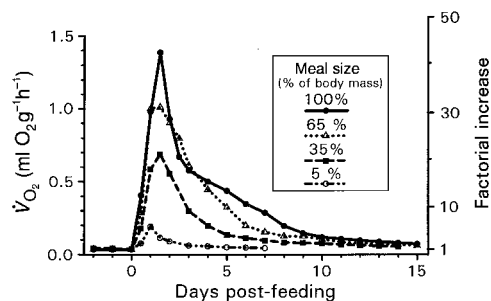


Figure 1 Oxygen consumption rates (left ordinate: $\dot{V}_{O_2}$ in units of $ml\ g^{-1}\ h^{-1}$) of juvenile Burmese pythons before and after consuming rodent meals equivalent to 5, 35, 65 and 100% of the snake's body mass. The right ordinate replots values as multiples of the fasting metabolic rate measured before feeding. Note that $\dot{V}_{O_2}$ rises steeply to a peak within 1–2 d of feeding, and then declines back to fasting levels within 4–16 d; and that larger meals elicit bigger and longer responses. For comparison, the factorial increase in $\dot{V}_{O_2}$ is up to 1.5 in a digesting human, 18 in a running human, and 45 in a galloping racehorse. Data are from ref. 4.

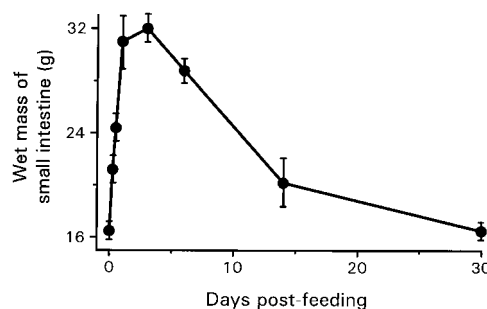


Figure 2 Wet mass of small intestine of pythons as a function of time after consuming rodent meals equivalent to 25% of snake body mass. Note that intestinal mass doubles within 24 h; dry mass yields essentially the same results. Data are from ref. 7.

by maintaining organs and proteins needlessly in readiness during pythons' characteristic long fasts? This qualitative reasoning, based on trade-offs of costs, exemplifies the general economic hypothesis regarding the evolution of regulation that we formulated earlier. The challenge in evaluating this qualitative hypothesis is to test it quantitatively, by actually measuring the costs being traded off against each other.

Snakes lend themselves to such a quantitative test. Many snake species besides pythons share pythons' feeding habits: expending little energy in hunting, waiting motionlessly to ambush prey, and then consuming large prey at infrequent intervals. But other snake species feed like humans and rats: they actively search for prey and consume small meals at frequent intervals. Hence we measured metabolic costs and regulatory spans of digestion in four species of infrequently feeding, ambush-hunting snakes and in four species of frequently feeding, actively foraging snakes, to all of which we fed rodent meals equivalent to 25% of the snake's body mass.

As summarized in Fig. 4, we found a clear physiological dichotomy between the two groups of snakes. (Phylogenetic independent-contrast analysis^{11,12} demonstrates that the dichotomy is indeed related to feeding habits and not to phylogenetic history.) Frequent feeders exceed infrequent feeders in standard metabolic rate by an average factor of 2.1. However, infrequent feeders exceed frequent feeders by a factor of 2.0 in relative cost of digestion (SDA coefficient); the coefficient's values of only 14–15% for frequently feeding snakes (much lower than the values of 21–35% for infrequent feeders) are similar to values for frequently feeding mammal, bird and invertebrate species⁴. Infrequent feeders also exceed frequent feeders by a factor of 2.4 for the factorial increase in $\dot{V}_{O_2}$ upon feeding. In none of the four frequent feeders does the intestinal brush-border uptake rate of any of the five studied solutes, nor the wet or dry mass of any of the eight studied internal organs, increase significantly on feeding. In all four infrequent feeders, however, the uptake rates of all five solutes are upregulated upon feeding (by factors of up to 28), as are the masses of the small intestine and liver in all four species and of the stomach, kidneys and pancreas in some species. That is, even if the attention is confined to snakes, infrequent feeders save on maintenance costs, as reflected in their lower standard metabolic rates (because they do not keep organs and proteins ready for the next meal), but they thereby incur larger start-up costs and require high regulatory spans with each meal.

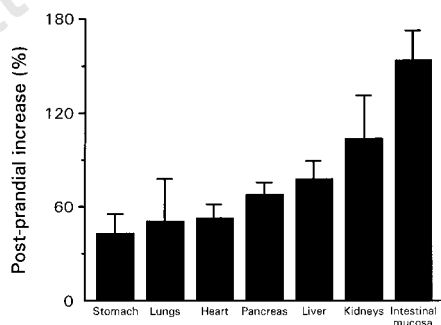


Figure 3 Percentage increase (as percentage of fasted mass) in the wet mass of python organs at 1 or 3 d after consuming rodent meals equivalent to 25 or 65% of snake body mass. Organ dry masses yield essentially the same results. Note that all organs increase rapidly in mass upon feeding, by at least 50%. Data from ref. 8.

Of these eight species of snake, the two for which we have the most information about feeding habits in the wild are the frequently feeding coachwhip (*Masticophis flagellum*) and the infrequently feeding sidewinder (*Crotalus cerastes*)^{1,13}. The former consumes meals averaging 15% of its body mass at 10-day intervals; the latter consumes meals averaging 25% of its body mass at 6-week intervals. Using physiological parameters that we measured for each species—standard metabolic rate and costs of digestion (SDA)—we calculated the sum of these two energy costs as a function of meal interval for each species. As illustrated in Fig. 5, the combined costs, which contribute to a large fraction of the snakes' energy budgets in the wild¹, cross over at a feeding interval of 4 weeks. At intervals of less than 4 weeks, a snake with sidewinder-like physiology (that is, large regulatory responses and low standard metabolic rate¹⁴) is thereby obliged to expend more energy (because it would often incur its high start-up costs), but it would incur lower expenditure for intervals longer than 4 weeks (because it saves on the coachwhip's high maintenance costs and rarely incurs its own high start-up costs). This calculation agrees with field observations: the coachwhip has co-evolved its physiology with a 10-day natural feeding interval, whereas the sidewinder's physiology has co-evolved with a 6-week natural feeding interval.

Thus, Fig. 5 confirms quantitatively for snake metabolic physiology the hypothesis that cost trade-offs caused regulation to evolve for some biological processes but not for others.

Outlook

We conclude by mentioning three fruitful directions for future research.

Cost trade-offs underlying the evolution of other regulatory systems. Figure 5 may serve as a prototype for calculations comparing

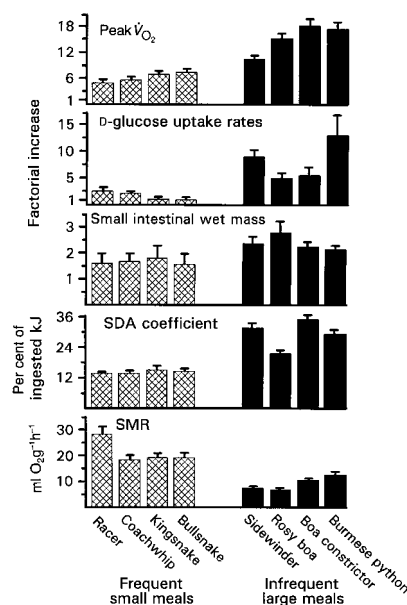


Figure 4 Factorial increases (relative to values for fasting individuals of the same species) of five parameters in four frequently feeding (left) and four infrequently feeding (right) snake species. The parameters are: whole-animal peak $\dot{V}_{O_2}$ consumption rate ($\dot{V}_{O_2}$), uptake rate by the small intestinal brush-border of D-glucose, small intestinal wet mass at 1 d post-feeding, and relative cost of digestion (SDA coefficient), all measured during digestion of a meal equivalent to 25% of the snake's body mass; and fasting metabolic rate (equivalent to standard metabolic rate or SMR, allometrically corrected to a body mass of 350 g). Note that the species that habitually consume infrequent large meals exhibit much bigger responses to feeding than do the species that habitually consume frequent small meals, even though all eight species were studied after consuming the same size of meal.

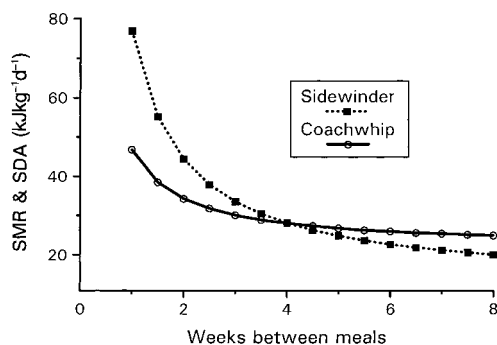


Figure 5 Time-averaged daily partial energy budgets as a function of interval between meals, calculated for two snake species that differ in physiology and in feeding ecology. Wild sidewinders capture prey that average 25% of their body mass at average intervals of 6 weeks. Wild coachwhips capture prey that average 15% of their body mass at intervals of 10 d. Energy budgets are calculated from fasting metabolic rates (SMR, standard metabolic rate) and costs of digestion (SDA) measured for each species, using average meal sizes in the wild but using various feeding intervals ranging from 1 to 8 weeks. Fasting metabolic rate is 2.1-fold higher in coachwhips, but the cost of digestion for equivalent meal sizes is 2.2-fold higher in sidewinders. Note that sidewinders have more a costly energy budget at feeding intervals less than 4 weeks but have a less costly budget at feeding intervals greater than 4 weeks.

the costs of other regulated and non-regulated systems, as a function of interval between demands on the system. Does this simple evolutionary reasoning account for why regulation has or has not evolved in other cases? In snakes, we could calculate the costs of digestion and of maintaining the underlying metabolic machinery from the whole snake's $\dot{V}_{O_2}$, because digestion and its machinery account for much of the snake's energy budget (20–40%)¹. In the case of regulated systems accounting for only a small fraction of the animal's energy budget, our proposed test will require somehow measuring or calculating in isolation the relative costs of upregulating and of maintaining the system. For instance, one could compare the cost of maintaining a high activity of a regulated protein, given constant protein turnover, with the cost of maintaining a low activity combined with infrequent bouts of increased synthesis of the same protein if it were regulated.

Other big eaters. By standards familiar to us (based on humans and rats), the metabolic physiology of ambush-hunting snakes is exceptional. Actually, many other animal species resemble them in alternating large meals with long fasts. These other species should also be examined for the possible existence of extreme metabolic regulation. The long list of candidate species includes lions, wolves, Komodo dragons and some abyssal fishes, which all consume huge meals; hibernating animals; birds and whales that migrate long distances without feeding; young seals, penguins and petrels that fast for several months after weaning or fledging before going to sea to feed; and adult penguins and albatrosses that fast for several months during courtship and egg incubation.

Pythons as model species in vertebrate regulatory biology. The large regulatory responses of pythons compared with humans and rats (Table 1) recommend them as a model species. In addition, they offer other advantages besides that quantitative one. Being vertebrates, their proteins possess much higher sequence homology with mammalian proteins than do squid and *Drosophila* proteins. The python's linear anatomy is convenient for surgical intervention, such as pancreatectomy or intestinal resection and reanastomosis. Contrary to the public image of pythons as being vicious and dangerous, juvenile pythons are docile and much less likely to bite than are rats, and are popular pets for children. They are cooperative experimental subjects, for instance in tolerating a swallowed intra-gastric pH electrode for one week. They are commercially available

through reptile breeders, are not an endangered species subject to legal restrictions, and are far cheaper, easier and cleaner to house, maintain and feed than are similarly sized mammals (only two meals and two combined semi-solid defecations/urinations per month). Their clutches of up to 100 eggs permit reducing inter-individual genetic variation by using siblings. They do not arouse the controversy associated with medical research on similarly sized mammals.

As examples of the potential research value of pythons, listed below are five projects that are currently proving profitable. (1) Molecular mechanisms of regulation of the glucose transporter SGLT1 are being studied in pythons over a much greater regulatory span (40-fold) than in rats (twofold) (M. Martín, E. Wright and S.M.S., unpublished observations). (2) Conlon *et al.*^{15,16} have isolated and sequenced eight python gastrointestinal hormones for synthesis and infusion to determine their physiological actions, and for developing assays to measure physiological release and plasma levels. Post-feeding surges in plasma levels prove to be high (for example, they increase by a factor of 52 for python cholecystokinin). Python insulin has an amino-acid sequence otherwise undocumented in nature, but previously developed by synthesis as a superactive analogue of human insulin¹⁵. (3) Surgically modified pythons with bypassed pancreas and biliary ducts, a balloon catheter in the stomach, surgically isolated intestinal loops, and chronically implanted intestinal catheters for infusing nutrient solutions, lend themselves to experiments designed to unravel neurohormonal regulatory signals. (4) The large and rapid organ growth that occurs in pythons within one day of feeding (Figs 2, 3) makes them convenient models for studying intestinal and kidney hypertrophy, and unique models for studying physiological pancreatic and cardiac hypertrophy (for which we have no mammalian model). (5) The very large increases in whole-animal $\dot{V}_{O_2}$, ventilation, heart rate, and blood flow of fed pythons make them convenient for studying cardiopulmonary regulation.

These examples illustrate the potential of juvenile pythons to become the equivalent of the squid axon in vertebrate regulatory biology. □

Stephen M. Secor and Jared Diamond are in the Department of Physiology, University of California Medical School, Los Angeles, California 90095-1751, USA.

1. Secor, S. M. & Nagy, K. A. Bioenergetic correlates of foraging mode for the snakes *Crotalus cerastes* and *Masticophis flagellum*. *Ecology* **75**, 1600–1614 (1994).
2. Greene, H. W. *Snakes: The Evolution of Mystery in Nature* (California Press, Berkeley, 1997).
3. Pope, C. H. *The Giant Snakes* (Knopf, New York, 1961).
4. Secor, S. M. & Diamond, J. Determinants of post-feeding metabolic response in Burmese pythons (*Python molurus*). *Physiol. Zool.* **70**, 202–212 (1997).
5. Diamond, P., Brondel, L. & LeBlanc, J. Palatability and postprandial thermogenesis in dogs. *Am. J. Physiol.* **248**, E75–E79 (1985).
6. Kleiber, M. *The Fire of Life: An Introduction to Animal Energetics* (Wiley, New York, 1961).
7. Secor, S. M. & Diamond, J. Adaptive response to feeding in Burmese pythons, pay before pumping. *J. Exp. Biol.* **198**, 1313–1325 (1995).
8. Secor, S. M. & Diamond, J. Effects of meal size on postprandial responses in juvenile Burmese pythons (*Python molurus*). *Am. J. Physiol.* **272**, R902–R912 (1997).
9. Jones, D. N. & Chappell, M. A. The effect of ration size and body size on specific dynamic action in Adélie penguin chicks, *Pygoscelis adeliae*. *Physiol. Zool.* **68**, 1029–1044 (1995).
10. Field, J., Belding, H. S. & Martin, A. W. An analysis of the relation between basal metabolism and summated tissue respiration in the rat. *J. Cell. Comp. Physiol.* **14**, 143–157 (1939).
11. Felsenstein, J. Phylogenies and the comparative method. *Am. Nat.* **125**, 1–15 (1985).
12. Garland, T., Harvey, P. H. & Ives, A. R. Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Syst. Biol.* **41**, 18–32 (1992).
13. Secor, S. M. Ecological aspects of foraging mode for the snakes *Crotalus cerastes* and *Masticophis flagellum*. *Herp. Monogr.* **9**, 169–186 (1995).
14. Secor, S. M., Stein, E. D. & Diamond, J. Rapid up-regulation of snake intestine in response to feeding: a new model of intestinal adaptation. *Am. J. Physiol.* **266**, G695–G705 (1994).
15. Conlon, J. M., Secor, S. M., Adrian, T. E., Mynarcik, D. C. & Whittaker, J. Purification and characterization of islet hormones (insulin, glucagon, pancreatic polypeptide and somatostatin) from the Burmese python, *Python molurus*. *Regul. Pept.* **71**, 191–198 (1997).
16. Conlon, J. M., Adrian, T. E. & Secor, S. M. Tachykinins (substance P, neurokinin A and neurokinin B), and neurotensin from the intestine of the Burmese python, *Python molurus*. *Peptides* **18**, 1505–1510 (1997).

Acknowledgements. This work was supported by grants from the NIH. We gratefully acknowledge our many collaborators named in our previous publications.

Correspondence should be addressed to S.M.S. (e-mail: python@ucla.edu).

Radio emission from the unusual supernova 1998bw and its association with the γ -ray burst of 25 April 1998

S. R. Kulkarni*, D. A. Frail†, M. H. Wieringa‡, R. D. Ekers§, E. M. Sadler||, R. M. Wark‡, J. L. Higdon‡, E. S. Phinney‡ & J. S. Bloom*

* Division of Physics, Mathematics & Astronomy, Caltech 105-24, Pasadena, California 91125, USA

† National Radio Astronomy Observatory, Socorro, PO Box 0, New Mexico 87801, USA

‡ Australia Telescope National Facility, CSIRO, Locked Bag 194, Narrabri, New South Wales 2390, Australia

§ Australia Telescope National Facility, CSIRO, PO Box 76, Epping, New South Wales 1710, Australia

|| School of Physics, University of Sydney, New South Wales 2006, Australia

¶ Division of Physics, Mathematics & Astronomy, Caltech 130-33, Pasadena, California 91125, USA

Data accumulated over the past year strongly favour the idea that γ -ray bursts lie at cosmological distances, although the nature of the power source remains unclear. Here we report radio observations of the supernova SN1998bw, which exploded at about the same time, and in about the same direction, as the γ -ray burst GRB980425. At its peak, the supernova was unusually luminous at radio wavelengths. A simple interpretation of the data requires that the source expanded with an apparent velocity of at least twice the speed of light, indicating that the supernova was accompanied by a shock wave moving at relativistic speeds (the ejecta of supernovae are typically characterized by non-relativistic velocities). The energy of the shock is at least 10^{49} erg, with an inferred ejecta mass of 10^{-5} solar masses, and we suggest that the early phase of this shock wave produced the burst of γ -rays. Although in general the properties of supernovae are very different from those of γ -ray bursts, we argue that this unusual supernova establishes a second class of γ -ray burst, which is distinctly different from the cosmological kind.

It is only within the past year that we have begun to understand γ -ray bursts (GRBs). At least one GRB is known definitely to be of extragalactic origin¹ and the inferred isotropic energy release, E_0 , of two others (ref. 2, and S. G. Djorgovski *et al.*, manuscript in preparation) has been estimated to be $\geq 10^{53}$ erg. This significant advance in our understanding is due to accurate localization of GRBs by the Italian–Dutch satellite BeppoSAX² and the discovery of relatively long-lived emission at X-ray³, optical⁴ and radio wavelengths⁵—the so-called ‘afterglow’ phenomenon. The burst itself and especially the afterglow emission are well accounted for by ‘fireball’ models^{6–9} which are similar to supernova models but with material (‘ejecta’) moving at relativistic speeds. The inferred mass of the ejecta (M_{ej}) is 10^{-5} solar masses, five to six orders of magnitude smaller than the mass of the ejecta in supernovae. Speculations about the cause of this explosion abound, but most involve the formation of a black hole; see refs 10, 11 for overviews.

Like GRBs, supernovae are also explosive events, but are much more numerous. The observed light curves and spectra are used to classify supernovae and, although it is not firmly established, astronomers believe that supernovae of type II, Ib and Ic mark the death of a massive star, resulting in the formation of a neutron star or a black hole. Supernovae of type Ia are popularly attributed to the destruction of a massive white-dwarf star. Most of the observable energy is initially in the form of the kinetic energy of the ejected mass. Estimates of the total kinetic energy E_0 for supernovae appear to cluster in a narrow range around 10^{51} erg.

The primary difference between supernovae and GRBs lies in the mass and hence in the speed and optical depth of the ejecta: high optical depth and non-relativistic speeds βc in supernovae (where c is the velocity of light), $\beta \leq 0.2$ or a bulk Lorentz factor, $\Gamma \leq 1.02$; and low optical depth and relativistic speeds in GRBs with bulk Lorentz factor $\Gamma \approx 300$. Here, following the standard convention,

$\Gamma \equiv (1 - \beta^2)^{-1/2}$. This difference accounts for the efficient high-energy emission of GRBs and the lower efficiency and lower energy (optical instead of γ -rays) emission of supernovae. Secondly, although the total energy E_T released is about the same ($E_T \approx 10^{53.5}$ erg) in supernovae (excluding Ia events) and GRBs, in normal supernovae, 99% of that energy is carried off by neutrinos. And the high optical depth means that 99% of the remaining 1% is converted into kinetic energy during adiabatic expansion. At interstellar particle densities $< 10 \text{ cm}^{-3}$, only 10^{-4} of E_T escapes in electromagnetic emission during the early months, the rest of the kinetic energy being radiated by the supernova remnant over the following $\sim 10^6$ yr.

Here we present radio observations of SN1998bw (ref. 12), the analysis of which leads us to suggest that this supernova is a link between some GRBs and supernovae. SN1998bw appears to have exploded within the joint time and spatial error box¹³ of GRB980425. The low probability (10^{-4}) of finding such a young supernova within the compact error circle of the GRB suggests an association¹² between SN1998bw and GRB980425. As discussed in detail below, the radio emission from SN1998bw has several unusual properties, including extraordinarily luminosity¹⁴. These unique features further strengthen¹⁵ the case for the association.

The arguments listed above in favour of the GRB-supernova association are compelling, but fail to provide an observational clue as to how a supernova can generate a burst of γ -rays. Theoretical models^{16,17} constructed to explain the optical observations of SN1998bw are simply unable to explain the observed burst of γ -rays. Below we show how a simple interpretation of the radio data forces us to conclude that SN1998bw had a shock wave moving at relativistic speed, ahead of the low-velocity ejecta that powers the optical light curve. If our interpretation is correct (and fortunately it is amenable to observational verification), then we have identified a

specific mechanism—a relativistic shock—that could potentially generate a burst of γ -rays at early times in this supernova.

GRB980425

Soffitta *et al.*¹³ detected a γ -ray burst of ~ 30 seconds duration in the BeppoSAX Gamma-Ray Burst Monitor and the Wide Field Camera (WFC) on 1998 April 25.91 UT. Follow-up observations^{18,19} of the 8-arcmin (radius) WFC error circle were carried out with the Narrow Field Instruments (NFI), also on-board BeppoSAX. Two faint X-ray sources were detected, of which one, 1SAX J1935.3–5252, faded between the two epochs, whereas the other remained steady. On April 28.73, we initiated our programme of radio observations at the Australia Telescope Compact Array (ATCA), an interferometric east–west array operated by the Australia Telescope National Facility.

We reported²⁰ the absence of radio sources within the 1-arcmin-radius error circles of the two NFI sources with typical $3\text{-}\sigma$ upper limits of 0.26 mJy and 0.3 mJy in the 3-cm and 6-cm bands, respectively. Similar limits were obtained with data obtained at later times; see Table 1 for a log of ATCA observations. Likewise, at optical wavelengths, Galama *et al.*²¹ and Bloom *et al.*²² reported the absence of any transient source brighter than R -band magnitude $R < 21$ with changes larger than ± 0.2 mag. The radio and optical upper limits are not sensitive enough to constrain the range of afterglow emission seen from γ -ray bursts located at cosmological distances.

SN1998bw

Inside the error circles of the WFC, but not coincident with either of the two NFI X-ray sources, Galama *et al.*²³ noted a previously unknown bright optical object seemingly located on the western spiral arm of the barred spiral galaxy ESO184-G82. Consequently, these authors suggested that the source was a possible supernova. In the course of searching for radio sources in the WFC error circle, we noted²⁰ that the brightest radio source in the WFC error circle, J193503.3–525045, coincided with the supernova candidate of

Galama *et al.* and that it too brightened considerably by May 5.6 UT. This strong early detection of a supernova candidate motivated us to begin a radio monitoring programme at several centimetre wave bands (Table 1). A summary of the various localizations can be found in Fig. 1.

From optical spectroscopic observations^{24,25} the redshift to ESO184-G82 is measured to be 0.0083 (heliocentric). The detection of narrow line absorption from sodium D lines (R. A. Stathakis *et al.*, manuscript in preparation) at the redshift of ESO184-G82 shows that the supernova is either in or behind ESO184-G82. The distance to this galaxy, assuming a Hubble constant of $65 \text{ km s}^{-1} \text{ Mpc}^{-1}$, is 38 Mpc.

Spectroscopic observations^{17,24,26} and their interpretation¹⁵ confirm that the object is a supernova, and accordingly the Central Bureau for Astronomical Telegrams designated this object as SN1998bw. The resemblance of the spectrum of this object to the usual supernova taxonomic categories is confusing. However, there is wide agreement that this is not a type Ia nor a type II supernova, but is a peculiar type Ib/c supernova. Furthermore, there are indications of higher-than-normal ejection speeds (approaching $60,000 \text{ km s}^{-1}$; R. A. Stathakis, personal communication; also ref. 18). As an aside we note that Iwamoto *et al.*¹² remark that it is unusually luminous for a type Ib/c supernova. However, there is considerable scatter²⁷ in the peak luminosity of type Ib/c: indeed, the type Ic SN1992ar was a magnitude brighter²⁸ than SN1998bw.

Radio observations of SN1998bw

The results of our radio monitoring programme are summarized in Table 1 and Fig. 2. The rapid early rise in the flux allows us to establish empirically the epoch of origin for the radio emission to better precision than that obtained from the optical measurements¹². The radio epoch is close (± 2 days) to the time of the γ -ray burst, thereby further strengthening the case for a physical association between the GRB and the supernova.

The observed radio properties allow us to make several statements about this supernova; we refer the reader to refs 29 and 30

Table 1 Radio flux density measurements of SN1998bw

Date (UT)	Elapsed time (d)	S_{20} (mJy)	S_{13} (mJy)	S_6 (mJy)	S_3 (mJy)	Array config.*	Int. time (h)
1998 Apr. 28	3.0			9.0	13.0	750A	3
1998 Apr. 29	4.0			9.9	13.0	750A	2
1998 May 05	9.9			39.0	48.0	750A	9
1998 May 07†	11.7	6.2	19.7	44.6	49.4	750A	3
1998 May 10	14.6	7.7	22.3	39.9	37.6	750A	2
1998 May 11	15.7	9.2	23.5	37.4	34.3	750A	2
1998 May 12	16.5	8.9	23.9	37.1	31.4	750A	1
1998 May 13	17.8	11.0	25.1	32.6	26.2	6C	3
1998 May 15	19.7	12.1	25.3	28.6	21.6	6C	0.5
1998 May 17	21.6	12.7	20.9	24.3	18.8	6C	1
1998 May 19	23.6	11.8	22.9	24.7	17.6	6C	1
1998 May 21	25.9	16.7	28.0	27.6	20.9	6C	0.5
1998 May 22	26.8	15.8	28.7	29.5	21.7	6C	0.7
1998 May 24‡	28.8	19.6	31.1	30.0	22.0	6C	1.3
1998 May 25‡	30.0	20.0	31.3	30.0	22.1	6C	2
1998 May 28	32.9	23.7	27.3	30.3	21.3	750E	1
1998 May 30	34.7	23.9	33.5	28.6	20.2	750E	1
1998 Jun 01	36.8	23.5	31.8	27.0	18.4	750E	1
1998 Jun 03	38.8	25.2	31.0	24.6	16.1	750E	0.8
1998 Jun 04	40.0	25.9	31.3	24.1	16.6	750E	0.6
1998 Jun 10	45.7	28.9	26.8	20.7	13.2	750E	1
1998 Jun 16	51.7	25.8	23.1	16.3	10.5	750E	2
1998 Jun 22	57.7	19.7	18.5	14.0	8.1	750E	2
1998 Jun 29	64.7	17.4	15.6	11.4	7.7	750E	3.8
1998 Jul 02	67.7	17.4	15.6	11.1	7.2	750E	2.3
1998 Jul 15	80.5	13.0	9.6	6.3	4.1	750E	1.7

Column headings (left to right) show: the UT date of the observation, the time since the start of the supernova explosion (calculated assuming that it occurred when the γ -rays from GRB980425 were first detected on 1998 March 25.90915 UT (ref. 14)), the flux density at 20 cm (1.38 GHz), 13 cm (2.49 GHz), 6 cm (4.80 GHz), and 3 cm (8.64 GHz), the array configuration for the ATCA, and the total integration time obtained for each pair of wavelengths (20 cm/13 cm and 6 cm/3 cm).

* All observations used a bandwidth of 128 MHz and two orthogonal linear polarizations for each wavelength pair. The individual antenna elements were moved in the course of this monitoring effort, forming different array configurations (750A, 6C, 750E). We minimized the effects of confusion, an important issue for compact configurations such as 750A and 750E, by subtracting background sources from the visibilities, and by excluding the shortest baselines from the analysis.

† The initial pointing centre was toward 1SAX J1935.3–5252 but on and after 1998 May 07 it was shifted to SN1998bw. All the tabulated flux density measurements of SN1998bw before this date were corrected by the primary beam response of the antennas.

‡ A search was made for linear or circularly polarized flux on day 12. No signal was detected in the Stokes Q , U and V above 0.5% and 0.9% at 3 cm and 6 cm, respectively. The limits on day 29 and 30 are $<1.5\%$ polarization at 3, 6 and 13 cm and $<2\%$ linear polarization at 20 cm.

and reviews of radio supernovae. First, SN 1998bw is not a type Ia supernova as no radio emission has ever been detected from a type Ia. Second, we note that the light curve has two main components: the first component peaks at about day 12, and the second component has a broad peak between days 30 and 40. Although it was some six orders of magnitude brighter, this two-component radio light curve is superficially similar to that of SN1987A (ref. 31). The rapid rise is not consistent with a normal type II radio supernova (RSN) but it is compatible with a type Ib/c RSN. Third, unlike the radio afterglow⁶ of GRBs, the light curves evolve quite smoothly. As discussed later, the smoothness of the light curve provides an important empirical constraint on the size of the source.

The 6-cm flux (S_6) peaks on day 12 and the spectral luminosity, $4\pi S_6 d^2$ is $\sim 8 \times 10^{28} d_{38}^2 \text{ erg s}^{-1} \text{ Hz}^{-1}$; here and below we assume that SN1998bw is located in the galaxy ESO184-G82 at a distance (d) of $38 d_{38}$ Mpc. This luminosity rivals that of the powerful RSN SN1998Z (ref. 34); in contrast to SN1998bw, this type II supernova achieved peak flux 3 years after the explosion. By day 40, it is clear that the flux in the shortest two wavebands (3 cm and 6 cm) are declining. The asymptotic spectral index (α) between these two bands is about -0.75 and can be reasonably interpreted as the spectral index of the optically thin portion of a synchrotron spectrum; here, the flux at frequency ν is assumed to be proportional to ν^α .

A single observation was also made on the night of 1998 May 7.7 UT, using the SCUBA bolometer array on the 15-m James Clark Maxwell Telescope (JCMT). Although SN1998bw was observed at low elevation ($<17^\circ$), it was detected in the 2-mm band with a flux density of 39 ± 11 mJy.

Constraint on the size of the RSN from variability

The expected angular size of the expanding radio photosphere θ_s can be estimated by assuming that the radio photosphere expands at

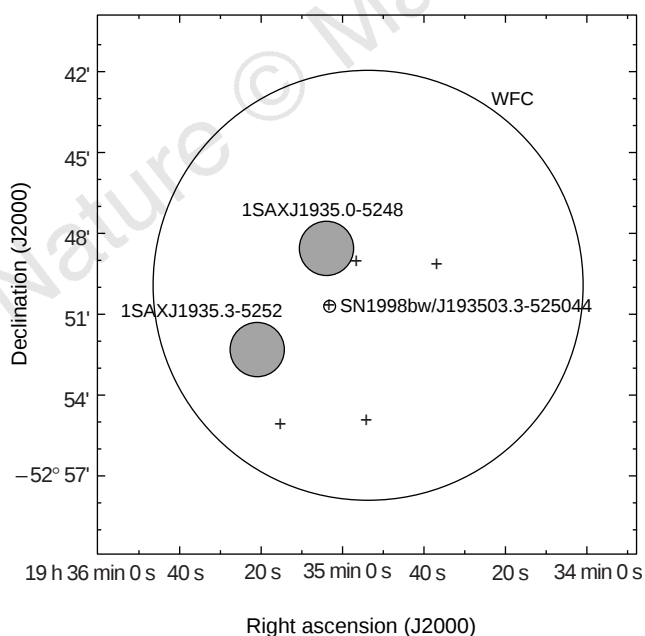


Figure 1 The position of objects within the 8-arcmin (radius) location error circle of the WFC for GRB980425 (ref. 13). The two NFI X-ray sources¹⁸ (the transient 1SAX J1935.3-5252 and the steady 1SAX J1935.0-5248) are indicated by shaded regions indicating their approximate position uncertainties. Small crosses indicate the positions of radio sources detected by the ATCA. The radio and optical emission from SN1998bw (J193503.3-525044) is indicated by a circle with a cross. The mean position for SN1998bw, obtained by averaging the best 3-cm and 6-cm observations, is right ascension (RA) 19 h 35 min 3.316 s and declination (dec) $-52^\circ 50' 44.75''$ (equinox J2000). The $1-\sigma$ error in RA is 0.01 s and $0.07''$ in dec.

the same speed as that inferred from optical spectroscopy which we noted earlier (R. A. Stathakis *et al.*, manuscript in preparation) to be $60,000 \text{ km s}^{-1}$. We obtain $\theta_s = 0.91 t_d v_{60} d_{38}^{-1} \mu\text{as}$, where $v_R = 6 \times 10^5 v_{60} \text{ km s}^{-1}$ is the expansion speed of the radio-emitting region and t_d is time since explosion in units of days. Thus the expected angular radius on day 10 is about $9 \mu\text{as}$. This is sufficiently small that one expects to see deep modulation of the received signal on timescales of hours due to scattering³³⁻³⁵ of the radio waves by density irregularities in the diffuse ionized medium of our Galaxy. It is worth noting that deep intensity variations have been seen towards radio afterglow of GRBs (see, for example, ref. 3) and the theory described below has been used to infer the angular size of GRBs.

In Fig. 3 we show the various regimes of interstellar scattering. We immediately note that diffractive scattering may have been important only at the earliest times. Unfortunately, during this time SN1998bw was not located at the field centre and antenna-pointing

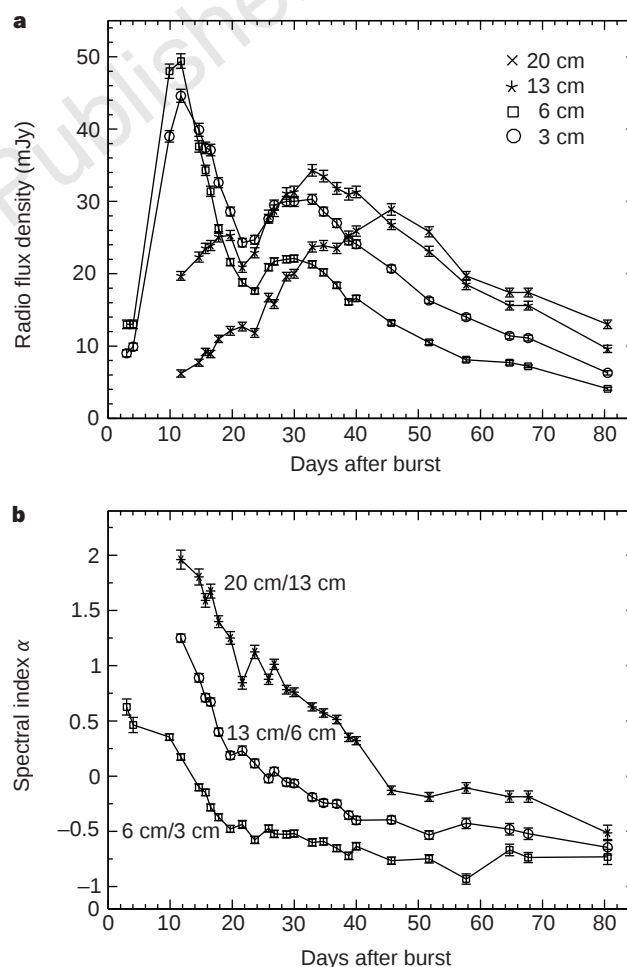


Figure 2 The radio light curve and spectrum of SN1998bw. **a**, The radio light curve at four wavelengths is shown. The age of the supernova has been calculated assuming that the explosion date can be given by the detection of γ -rays from GRB980425. The error bars given on the plot are larger than the formal errors estimated from the receiver noise owing to the difficulties of determining the flux density with short snapshots. The errors, including absolute flux scale errors, thermal noise and confusion noise, can be approximated by the quadrature sum of a constant term (0.5 mJy for both 20 cm and 13 cm, 0.3 mJy for 6 cm and 0.2 mJy for 3 cm) and a term that is a fraction of the flux density (2%). **b**, The evolution of the radio spectral index for SN1998bw. The spectral index α (where $S_\nu \propto \nu^\alpha$) is calculated between 20 cm and 13 cm (crosses), 13 cm and 6 cm (circles) and 6 cm and 3 cm (squares). The age of the supernova has been calculated assuming that the explosion date is given by the detection of γ -rays from GRB980425.

errors probably dominate the observed variations. Outside these times, we are in the weak-scattering or refractive-scattering regimes.

Using an upper bound of three times the observed fractional variation on day 12 (the measured variations were 6% at 6-cm and 3% at 3-cm wavelengths over a total duration of 7 hours) and the formulation given in refs 34, 35, we derive a lower bound on the velocity of expansion of the radio photosphere of 70,000–90,000 km s⁻¹. However, stronger constraints come from the lack of variability at lower frequencies. This is best seen in Fig. 3, in which we see the 20-cm and 13-cm observations lie squarely in the refractive regime. Beginning at day 12, the fluxes at 20-cm and 13-cm wavelengths rise smoothly over the next 10 days. After subtracting the smooth rise, we find day-to-day variations of 5%

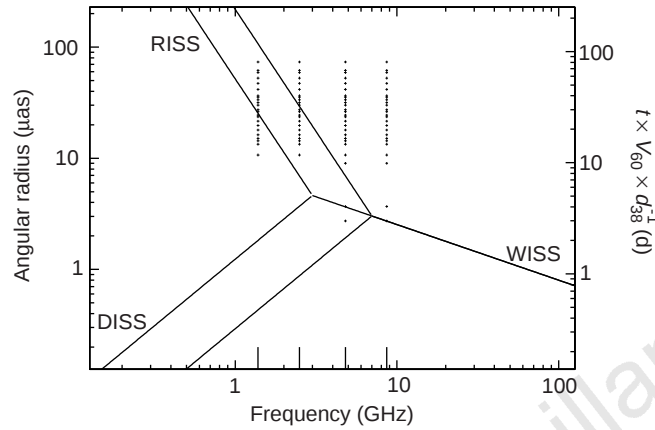


Figure 3 The different regimes of interstellar scattering. For the purpose of theoretical modelling, the Galactic ionized medium can be conveniently approximated by a thin screen located at an effective distance of $D \approx 0.5$ –1 kpc. Scattering is considered to be in the ‘weak’ regime when the Fresnel scale, $r_F = \sqrt{\lambda D/2\pi}$ is smaller than r_0 , the spatial scale of the irregularities in the scattering screen; here λ is the wavelength of observations. In this regime, the intensity variations are modest and m , the modulation index (defined as the ratio of the r.m.s. of the intensity variations to the mean) is well below unity. When $r_0 \ll r_F$, the scattering is considered to be in the ‘strong’ regime and the multitude of r_0 -sized patches within a single Fresnel scale results in multi-path propagation. This regime splits into two regimes: diffractive and refractive, both of which exhibit deep intensity variations but over a narrow band and a broad band, respectively. Both r_0 and r_F are functions of the observing frequency. Instead of talking in terms of r_0 and r_F it is more convenient to talk in terms of the transition frequency³⁵, ν_0 . Observations at frequency (ν) above ν_0 are in the weak regime, and observations at lower frequency suffer from strong scattering. Assuming typical interstellar-medium parameters for this direction, ν_0 is in the frequency range 3–7 GHz (ref. 35), a frequency range straddled by our ATCA observations. The lines in this figure indicate the frequency dependence of diffractive (DISS), refractive (RISS) and weak (WISS) scattering as a function of the source angular radius (left-hand axis) for the above range of ν_0 values. The Fresnel angle, $\theta_F \equiv r_F/D$ evaluated at the transition frequency of (say) 5 GHz is $\theta_F \approx 4 \mu\text{as}$; here and below we have set D to 1 kpc. Interstellar scattering allows us to constrain source size as long as the source is smaller than one of three characteristic angular scales: (1) ‘weak’, $\theta_W = \theta_0(\nu_0/\nu)^{1/2}$, (2) ‘refractive’, $\theta_R = \theta_0(\nu_0/\nu)^{1/5}$ and (3) ‘diffractive’, $\theta_D = \theta_0(\nu_0/\nu)^{-6/5}$. On the right-hand axis we plot the expected time (in days) that SN1998bw would reach this radius if the velocity of the radio photosphere in units of 60,000 km s⁻¹ (v_{60}) is 1 and the source distance in units of 38 Mpc (d_{38}) is 1. Bold ticks at the bottom of the figure indicate the frequencies used in the ATCA observations: 1.38 GHz (20 cm), 2.49 GHz (13 cm), 4.8 GHz (6 cm), and 8.6 GHz (3 cm). The small crosses are plotted at the frequencies and dates at which ATCA observations were made. We note that on day 10, the 6-cm flux shows a very smooth increase over the observing interval of 9 h, whereas the 3-cm flux increased but with large scatter (20%). We believe that this scatter could be primarily due to pointing errors, as SN1998bw was located on the half-power response of the 3-cm primary beam for this observation.

at 20-cm and 1% at 13-cm wavelengths (over a period of 7 days). Comparing these variations to the expected daily value of $\sim 35\%$ (ref. 34) leads us to conclude that the radio expansion speed is at least $0.3c$.

Modelling the radio light curves

Radio observations of supernovae have been interpreted in the framework of the ‘mini-shell’ model³⁶. The ejecta push a decelerating shock ahead of themselves into the circumstellar material. The shocked circumstellar material is then subject to Rayleigh–Taylor instability. This instability can drive turbulent motions which amplify magnetic fields and accelerate particles to relativistic energies via the Fermi process³⁷. Accelerated electrons gyrate in the newly amplified magnetic field and generate strong synchrotron emission. It is generally assumed that the energy in the magnetic field is comparable to that in the energetic particles.

We note that there has been considerable observational support for the mini-shell model. We stress that this model (or the assumptions leading to this model, power-law distribution and magnetic fields in equipartition with the particles) forms the foundation of the interpretation given here.

The fact that the observed radio flux density drops at low frequency while the flux density of optically thin synchrotron radiation would continue to increase with decreasing frequency requires a frequency-dependent absorption mechanism. The two dominant mechanisms are free–free absorption by the surrounding circumstellar matter (assumed to be ionized by the supernova or the progenitor) and synchrotron self-absorption. Chevalier³⁶ concluded that the delayed turn-on of the lower-frequency radio emission is best explained by free–free absorption. However, starting with Shklovskii³⁸, there has been gradual recognition (refs 30, 39, and S.R.K. and E.S.P., manuscript in preparation) of the importance of synchrotron self-absorption in radio supernovae which exhibit early radio emission; that is type Ib/c radio supernovae.

We now return to SN1998bw. We noted that the radio emission is composed of two emission components (Fig. 2a). The first component, on about day 10, peaks in the 3-cm band with a flux of 50 mJy. This component has some of the features of a classical synchrotron self-absorbed spectrum. Specifically, on day 10 we infer a peak frequency, $\nu_p \approx 10$ GHz and associated peak flux $S_p \approx 50$ mJy. It is well known that in a homogeneous source with a power-law electron spectrum, the self-absorbed synchrotron spectrum exhibits a $\nu^{5/2}$ power law for frequencies less than ν_p . The observed spectral index of 2 between 20-cm and 3-cm wavelengths (Table 1; Fig. 2b) is in acceptable agreement with this expectation. Having made a reasonable case for a synchrotron self-absorbed spectrum, we now apply diagnostics of synchrotron theory to our data.

Two of us (S.R.K. and E.S.P., manuscript in preparation) have shown that two key diagnostics—the brightness temperature (T_B) and the equipartition radius⁴⁰—can be used draw robust conclusions from light curves of radio supernovae, in spite of the (unknown) complication of external absorption. We now apply these two diagnostics in turn. The brightness temperature is given by:

$$T_B(\nu) = \left(\frac{c^2}{2k} \right) \left(\frac{S(\nu)}{\pi \theta_s^2} \right) \nu^{-2} \quad (1)$$

where k is the Boltzmann constant. The value of this diagnostic is that it is a well known result^{41,42} that the T_B of a source radiating via the incoherent synchrotron mechanism is limited to $T_{\text{icc}} \approx 10^{12}$ K. The origin of this limiting T_B is the ‘inverse-Compton catastrophe’. When T_B exceeds T_{icc} , multiple inverse Compton scattering of the synchrotron radio photons to γ -ray/X-ray energies begins to rapidly dominate the luminosity.

For our specific case ($\alpha = 1$, $\nu_p = 5$ GHz), following Readhead⁴³, the ratio of the inverse Compton (L_{ic}) luminosity to the luminosity

in synchrotron photons (L_{synch}) is given by:

$$\frac{L_{\text{ic}}}{L_{\text{synch}}} = \left(\frac{T_B}{4 \times 10^{11} \text{ K}} \right)^5 \left[1 + \left(\frac{T_B}{4 \times 10^{11} \text{ K}} \right)^5 \right] \quad (2)$$

From the BeppoSAX NFI X-ray observations on day 1 and day 8, we infer an upper limit (3σ) on the 1.6–10 keV flux from the supernova, $f_X < 10^{-13} \text{ erg cm}^{-2} \text{ s}^{-1}$. On day 10, when the radio emission peaks in the centimetre-wave radio band, the 6-cm flux, $S_R = 50 \text{ mJy}$ and thus the flux in the radio band is $S_R \nu_R \approx 2.5 \times 10^{-15} \text{ erg cm}^{-2} \text{ s}^{-1}$. The ratio of the X-ray luminosity (if any) to the radio luminosity is thus below 40; using this value in equation (2), we obtain $T_{\text{ic}} \leq 5 \times 10^{11} \text{ K}$. A more precise value depends on the cut-offs in the particle distribution and the T_B adopted, which set the location of the ‘Compton humps’ in the spectrum. Indeed, a more severe constraint⁴⁴ comes from inverse Compton scattering of the optical photons (from the supernova) by the electrons.

The inferred brightness temperature is $T_B = 2.0 \times 10^{13} S(\text{mJy})(\lambda/6 \text{ cm})^{-2} \nu_{60}^{-2} t_d^{-2} d_{38}^2 \text{ K}$; here S is the flux at wavelength λ . Using the fluxes from Table 1, the estimated T_B exceed 10^{12} K for the first months and approach 10^{13} K in the first week. Clearly, we have erred in one of our assumptions. As discussed earlier, optical observations place a lower limit to the distance of SN1998bw and thus one cannot decrease T_B by assuming that SN1998bw is background to ESO184-G82. The origin of the time for the radio emission is reasonably secure (to within a day or so), as can be seen from looking at Fig. 2a. This leaves only two possibilities: either the assumed size of the source (that is, the assumed ν_R) is wrong or the inverse Compton arguments listed above are incorrect. Fortunately, as discussed by S.R.K. and E.S.P. (manuscript in preparation) within the frame of synchrotron theory, one can avoid the latter concern by considering the total energetics.

Briefly, the energetics argument is as follows. The energy in a synchrotron emitting source is in the form of relativistic electrons (U_e) and magnetic field strengths (U_B); there could be a substantial amount of energy locked in protons, but as protons do not radiate we cannot constrain this additional energy. It can be shown that $U_B \propto S_p^{-4} \nu_p^{10} r^{11}$ and $U_e \propto S_p^4 \nu_p^{-7} r^{-6}$, where r is the radius of the source. The total energy $U = U_e + U_B$ thus depends sensitively on r . Furthermore, one term (U_B) increases rapidly with r , whereas the other (U_e) decreases rapidly with r . Unfortunately, we have only a lower bound on r . The total energy is minimized when $r = r_{\text{eq}}$, the so-called ‘equipartition’ radius⁴⁰; the resulting total energy is the equipartition energy, U_{eq} .

It is convenient to express the total energy in terms of the equipartition energy;

$$\frac{U}{U_{\text{eq}}} = \frac{1}{2} \eta^{11} (1 + \eta^{-17}) \quad (3)$$

where $\eta = \theta_s/\theta_{\text{eq}}$ and:

$$\theta_{\text{eq}} = 120 \left(\frac{d}{\text{Mpc}} \right)^{-1/17} \left(\frac{S_p}{\text{mJy}} \right)^{8/17} \left(\frac{\nu_p}{\text{GHz}} \right)^{(-2\alpha - 35)/34} \mu\text{as} \quad (4)$$

We note that many authors (including ref. 40) define θ_{eq} to be a diameter and $S_p \propto \nu^{-\alpha}$, while we define θ_{eq} to be a radius and $S_p \propto \nu^\alpha$.

From equation (3) we see that a steep price in total energy U has to be paid if η deviates from unity. We now investigate the energetics. If the radio-emitting region expanded at the same speed as the optical shock, then $U \approx 10^{54} \text{ erg}$. If we allowed the radio shock to expand faster but insist that the resulting T_B not violate the X-ray inverse Compton constraint (see above), then the mean expansion speed would be $0.3c$ and $U \approx 10^{52} \text{ erg}$. In both cases, the inferred minimum energy exceeds the available energy of a typical supernova, 10^{51} erg . Thus we are forced to consider even more rapid expansion.

A relativistic shock

Above, using two separate lines of reasoning (lack of interstellar scintillation) and minimum energy arguments), we concluded that the radio shock must be expanding considerably faster than the shock producing the optical emission. We now estimate the shock speed at several epochs. To this end, it is far more convenient to frame the discussion in terms of brightness temperature rather than sizes. The requirement that θ_s must be comparable to θ_{eq} is equivalent to insisting that the inferred T_B must be comparable to the equipartition temperature⁴³, T_{eq} , which is obtained by substituting $\theta_s = \theta_{\text{eq}}$ in equation (1). For parameters used for day 16, $T_{\text{eq}} = 5 \times 10^{10} \text{ K}$. We note the following: first, T_{eq} depends very slightly on the observables, and for all practical purpose can be considered to be a constant; second, T_{eq} is almost an order of magnitude smaller than T_{ic} .

We now compute the brightness temperature of a relativistic expanding synchrotron source. The specific intensity is $I_\nu = (S/\pi)(d^2/(\Gamma\beta ct)^2) = 2kT_B \nu^2/c^2 = 2kDT'_B(\nu')(D\nu')^2/c^2$. Here the primed quantities are measured in the frame of the emitting material. The last equality follows because $I_\nu \nu^{-3}$ is a Lorentz invariant, and the Doppler factor $D = [\Gamma(1 - \beta \cos\theta)]^{-1}$ is of the order Γ for the fastest-moving material which dominates the emission. Thus the brightness temperature in the frame of the synchrotron-emitting plasma is:

$$T'_B = 8 \times 10^{11} \Gamma^{-3} S(\text{mJy})(\lambda/6 \text{ cm})^2 t_d^{-2} d_{38}^2 \text{ K} \quad (5)$$

On day 16, we assume $\nu_p = 6 \text{ GHz}$ and note that $S_p = 40 \text{ mJy}$. Insisting that $T'_B = T_{\text{eq}}$ we obtain: $\theta_{\text{eq}} = 100 \mu\text{as}$, $B \approx 0.4 \text{ G}$, and $U_{\text{eq}} = 10^{49} \text{ erg}$. On days 4, 29 and 60, we adopt $(\nu_p/\text{GHz}, S_p/\text{mJy}) = (10, 15), (3, 30)$ and $(1, 20)$, respectively. These yield angular sizes of 35, 160 and $400 \mu\text{as}$ and equipartition energies of $10^{48}, 2 \times 10^{49}$ and $1.5 \times 10^{49} \text{ erg}$. The corresponding mean expansion speeds are $r/t = 1.9c, 1.2c$ and $1.5c$. The transverse expansion speed is given by $\Gamma\beta c$ and thus the inferred Γ is between 1.6 and 2—the radio shock front is moving at relativistic speeds.

We note some caveats. (1) We have ignored possible external absorption (free–free absorption from the circumstellar material). If these are significant, then U_{eq} and the inferred mean speeds are strict lower bounds. (2) The tightest constraint on the inferred speed comes from the earliest observations for which ν_p is not well determined. However, for a simple self-absorbed synchrotron source, one can demonstrate that the use of a frequency other than ν_p again results in strict lower bounds. (3) We cannot account for the existence of the second peaked component in the radio light curve near day 30 (Fig. 2a), although it too requires relativistic expansion.

Perhaps the most serious concern is that our analysis is based on a static synchrotron source, whereas the source in question is expanding relativistically. Indeed, there is a suggestion that the broad-band spectrum on day 10 (when we have the widest spectral coverage) at high frequency does not resemble a classical synchrotron spectrum. Specifically, the spectrum is flat (zero spectral index) all the way from 6-cm to 2-mm wavelength.

Geometry and energetics

We concluded above that the radio emission in RSN 1998bw arises in a shock with an initial $\Gamma \approx 2$, which appears to be slowing down as the supernova aged. The minimum total energy in the radio-emitting region is $\sim 10^{49} \text{ erg}$, a rather impressive energy considering that most of the ejected mass is in a slower-moving shock which manifests itself via the optical emission.

We note that we have assumed spherical geometry. However, it could be that the radio emission is in the form of jets. This geometry has the advantage of reducing the energy requirements considerably. Fortunately, polarization measurements offer the possibility of probing the geometry of the emitting region. However, as can be

seen from Table 1, we find no evidence for either a circularly polarized or a linearly polarized signal.

The circular polarization of synchrotron radiation from relativistic particles with a smooth pitch-angle distribution is small unless the radiation is close to the cyclotron limit. The absence of circular polarization is thus not surprising, though it does place a constraint on very slow jet models with magnetic fields well above equipartition.

Shear and expansion in non-spherical synchrotron sources like jets generally create magnetic fields with enough anisotropy that their integrated polarizations are high. For example, the integrated fractional polarizations of radio quasars (including the core, the jet and the lobes) at 5 and 8 GHz are rarely below 4% (ref. 45), and bursts in BL Lac objects (believed to be caused by a single shock, as in the case of our supernova) have integrated linear polarizations of $\sim 10\%$ (refs 46, 47). The lack of any linearly polarized signal could be reconciled with the jet hypothesis if the jet made an angle much greater than Γ^{-1} with respect to our line of sight. Because of aberration, we would then be seeing the back of the shock almost face-on, so a tangled field in the shock plane would produce no net polarization. However, this hypothesis would require that the intrinsic Γ and luminosity of the jet are much larger than the ones inferred above. This seems implausible. We conclude that the simplest interpretation of the observed low polarization is that we are seeing a spherical blast wave.

A new type of GRB?

In our introduction we noted several circumstantial reasons favouring the association of GRB980425 and SN1998bw. If this association is correct, then GRB980425 is a new type of GRB, very different in its optical and radio properties and decidedly less energetic than the three GRBs at high redshift (refs 1, 2, and S. G. Djorgovski *et al.*, manuscript in preparation).

We now consider a possible origin for the observed burst of γ -rays. After day 4, the synchrotron lifetimes of the electrons in SN1998bw radiating at ~ 5 GHz are (for equipartition fields) ~ 1 yr. We note that the total energy radiated in the observed part of the radio spectrum is $\sim 10^{45}$ erg, much less than U_{eq} . But the flat spectrum implied by the flux in the 2-mm band suggests that higher-energy electrons may be more nearly radiative. The X-ray flux limit implies that the inverse Compton losses cannot be much greater, so the shock appears to be non-radiative throughout the observing period. Extrapolating back in time in a simple equipartition model, in which pressure $p \propto r^{-2}$ and the particle spectrum is independent of radius, the optically thin synchrotron luminosity from radius r to $2r$ scales as r^{-1} , as does the ratio of first inverse Compton to synchrotron luminosity. Thus at early times, a shock in the circumstellar gas or an internal shock in the ejecta would be radiative and predominantly a γ -ray source.

We therefore advance the hypothesis that in some supernovae the outer layer of the exploding star is given sufficient energy to cause it to expand relativistically. This initially produces a burst of γ -rays and subsequently is responsible for the radio emission. We note that the γ -ray burst energy of GRB980425 (at an assumed distance of 38 Mpc) is 10^{48} erg, which is smaller than U_{eq} . The mass in the relativistically expanding shock is $U_{\text{eq}}/(\Gamma - 1)c^2 \approx 10^{-5}$ solar masses, rather similar to that invoked in models⁷⁻⁹ for the cosmological GRBs.

More than two decades ago, Colgate⁴⁸ had proposed γ -ray emission from supernovae as a possible origin of GRBs. The proposed mechanism, acceleration of the supernova shock to relativistic speed down the density gradient of a massive stellar envelope, has not been confirmed by more detailed calculations, including radiation coupling (see, for example, ref. 49). Yet such a mechanism might work with the much smaller and less massive envelopes of carbon and helium stars appropriate to type Ib/c supernovae.

As with any advance, there are many significant open questions. How is the shock with the required Γ generated? Is the shock spherical or collimated? How much energy beyond the minimum is involved in the relativistic shock? Is the phenomenon common to all type Ic supernovae? Are such GRBs characterized by a special burst profile^{50,51}?

Recent work by Waxman and Loeb⁴⁴ offers a radical interpretation of the radio emission. In this "thermal" model, contrary to the basic tenets of the mini-shell model, the shock is subrelativistic and, more important, the shocked electrons rapidly thermalize with the protons and the magnetic fields are not in equipartition. In essence, if this model is right, we are exploring an entirely new phenomenon in shock physics, thanks to radio observations of SN1998bw.

Fortunately, these questions and outstanding issues can be answered by observations. Early-time observations, especially sub-millimetre observations, provide the best diagnostic of the fastest shock. For example, the 2-mm observation is incompatible with the thermal model of Waxman and Loeb⁴⁴. Late-time observations are equally important. In our model, the spectral index should be constant whereas curvature (as appears to be observed) is expected in the thermal model. Finally, we note that very-long-baseline interferometry (VLBI) observations could directly measure the geometry and size of the fast shock, and provide the most direct way to confirm or reject our proposed model. $\square$

Received 30 June; accepted 10 September 1998.

1. Metzger, M. R. Spectral constraints on the redshift of the optical counterpart to the γ -ray burst of 8 May 1997. *Nature* **387**, 878–880 (1997).
2. Kulkarni, S. R. *et al.* Identification of a host galaxy at redshift $z = 3.42$ for the γ -ray burst of 14 December 1997. *Nature* **393**, 35–39 (1998).
3. Boella, G. *et al.* BeppoSAX, the wide band mission for X-ray astronomy. *Astron. Astrophys. Suppl. Ser.* **122**, 299–399 (1997).
4. Costa, E. *et al.* Discovery of an X-ray afterglow associated with the γ -ray burst of 28 February 1997. *Nature* **387**, 783–785 (1997).
5. Van Paradijs, J. *et al.* Transient optical emission from the error box of the γ -ray burst of 28 February 1997. *Nature* **368**, 686–688 (1997).
6. Frail, D. A., Kulkarni, S. R., Nicastro, L., Feroci, M. & Taylor, G. B. The radio afterglow from the γ -ray burst of 8 May 1997. *Nature* **389**, 261–263 (1997).
7. Mészáros, P. & Rees, M. J. Optical and long-wavelength afterglow from gamma-ray bursts. *Astrophys. J.* **476**, 232–237 (1997).
8. Vietri, M. The afterglow of gamma-ray bursts: the cases of GRB 970228 and GRB 970508. *Astrophys. J.* **488**, L105–L108 (1997).
9. Waxman, E. Gamma-ray-burst afterglow: supporting the cosmological fireball model, constraining parameters, and making prediction. *Astrophys. J.* **485**, L5–L8 (1997).
10. Fishman, G. J. & Meegan, C. A. Gamma-ray bursts. *Annu. Rev. Astron. Astrophys.* **33**, 415–458 (1995).
11. Piran, T. in *Unsolved Problems in Astrophysics* (eds Bahcall, J. N. & Ostriker, J. P.) 343–377 (Princeton Univ. Press, 1997).
12. Galama, T. J. *et al.* An unusual supernova in the error box of the γ -ray burst of 25 April 1998. *Nature* **395**, 670–672 (1998).
13. Soffitta, P. *et al.* *IAU Circ.* No. 6884 (1998).
14. Wieringa, M. *et al.* *IAU Circ.* No. 6896 (1998).
15. Sadler, E. M., Stathakis, R. A., Boyle, B. J. & Ekers, R. D. *IAU Circ.* No. 6901 (1998).
16. Woosley, S. E., Eastman, R. G. & Schmidt, B. P. Gamma-ray bursts and Type Ic supernovae: SN 1998bw. Preprint astro-ph/9806299 available at (<http://xxx.lanl.gov>) (1998).
17. Iwamoto, K. *et al.* A hypernova model for the supernova associated with the γ -ray burst of 25 April 1998. *Nature* **395**, 672–674 (1998).
18. Pian, E. *et al.* *GCN Note*. No. 61 (1998).
19. Pian, E., Frontera, F., Antonelli, L. A. & Piro, L. Wieringa, M. *GCN Note*. No. 69 (1998).
20. Wieringa, M., Frail, D. A., Kulkarni, S. R., Higdon, J. L. & Wark, R. *GCN Note*. No. 63 (1998).
21. Galama, T. J. *et al.* *GCN Note*. No. 62 (1998).
22. Bloom, J. S., Kulkarni, S. R., Djorgovski, S. G., McCarthy, P. & Frail, D. *GCN Note*. No. 64 (1998).
23. Galama, T. J. *et al.* *IAU Circ.* No. 6895 (1998).
24. Lidman, C. *et al.* *IAU Circ.* No. 6895 (1998).
25. Tinney, C., Stathakis, R., Cannon, R. & Galama, T. *IAU Circ.* No. 6896 (1998).
26. Patat, F. & Piemonte, A. *IAU Circ.* No. 6918 (1998).
27. Miller, D. L. & Branch, D. Supernova absolute-magnitude distributions. *Astron. J.* **100**, 530–539 (1990).
28. Hamuy, M. *et al.* *IAU Circ.* No. 5574 (1992).
29. Weiler, K. W. & Sramek, R. A. Supernovae and supernova-remnants. *Rev. Astron. Astrophys.* **26**, 295–341 (1988).
30. Chevalier, R. A. Synchrotron self-absorption in radio supernovae. *Astrophys. J.* **499**, 810–819 (1998).
31. Ball, L., Campbell-Wilson, D., Crawford, D. F. & Turtle, A. J. Radio observations of SN 1987A at 843 MHz. *Astrophys. J.* **453**, 864–872 (1995).
32. van Dyk, S. D., Weiler, K. W., Sramek, R. A. & Panagia, N. SN 1988Z: the most distant radio supernova. *Astrophys. J.* **419**, L69–L72 (1993).
33. Narayan, R. The physics of pulsar scintillation. *Phil. Trans. R. Soc. Lond. A* **341**, 151–165 (1992).
34. Goodman, J. Radio scintillation of compact extra-galactic radio sources. *Mon. Not. R. Astron. Soc.* **294**, 307–311 (1998).
35. Walker, M. A. Interstellar scintillation of compact extra-galactic radio sources. *Mon. Not. R. Astron. Soc.* **294**, 307–311 (1998).
36. Chevalier, R. A. The radio and X-ray emission from type II supernovae. *Astrophys. J.* **259**, 302–310 (1982).
37. Ball, L. & Kirk, J. G. The acceleration of electrons in the radio supernova SN 1986J. *Astron. Astrophys.* **303**, L57–L60 (1995).

38. Shklovskii, I. S. Synchrotron self-absorption of the radio emission of supernova 1983.51. *Sov. Astron. Lett.* **11**, 105–106 (1985).
39. Slysh, V. I. Synchrotron self-absorption of radio emission from supernovae. *Sov. Astron. Lett.* **16**, 339–342 (1990).
40. Scott, M. A. & Readhead, A. C. S. The low-frequency structure of powerful radio sources and limits to departures from equipartition. *Mon. Not. R. Astron. Soc.* **180**, 539–550 (1977).
41. Hoyle, F., Burbidge, G. R. & Sargent, W. L. W. On the nature of the quasi-stellar sources. *Nature* **209**, 751–753 (1966).
42. Kellerman, K. I. & Pauliny-Toth, I. I. K. The spectra of opaque radio sources. *Astrophys. J.* **155**, L71–L78 (1968).
43. Readhead, A. C. S. Equipartition brightness temperature and the inverse Compton catastrophe. *J.* **426**, 51–59 (1994).
44. Waxman, E. & Loeb, A. A sub-relativistic shock model for the radio emission of SN 1998bw. Preprint astro-ph/9808135 available at (<http://xxx.lanl.gov>) (1998).
45. Conway, R. G. *et al.* The radio polarization of quasars. *Mon. Not. R. Astron. Soc.* **168**, 137–162 (1974).
46. Hughes, P. A., Aller, H. D. & Aller, M. F. Synchrotron emission from shocked relativistic jets. I. The theory of radio-wavelengths variability and its relation to superluminal motion. *Astrophys. J.* **341**, 54–67 (1989).
47. Hughes, P. A., Aller, H. D. & Aller, M. F. Synchrotron emission from shocked relativistic jets. I. A model for the centimeter wave band quiescent and burst emission from BL Lacertae. *Astrophys. J.* **341**, 68–79 (1989).
48. Colgate, S. A. Early gamma rays from supernovae. *Astrophys. J.* **187**, 333–335 (1974).
49. Ensmann, L. & Burrows, A. Shock breakout in SN 1987A. *Astrophys. J.* **393**, 742–755 (1992).
50. Bloom, J. S., Kulkarni, S. R., Harrison, F., Prince, T. & Frail, D. A. Expected characteristics of the subclass of supernova gamma-ray bursts. *Astrophys. J.* **506**, L105–L108 (1998).
51. Norris, J. P., Bonnell, J. T. & Watanabe, K. Constraints on association of single-pulse gamma-ray bursts and supernovae. Preprint astro-ph/9807322 available at (<http://xxx.lanl.gov>) (1998).

Acknowledgements. We thank L. Avery and G. Moriarty-Schieven for help in making the JCMT observations. D.A.F. and S.R.K. thank M. Rupen, M. Rees and B. Paczynski for discussions. S.R.K. thanks A. Readhead for extensive discussions of brightness temperature. The Australia Telescope is funded by the Commonwealth of Australia for operation as a National Facility managed by CSIRO. The James Clerk Maxwell Telescope is operated by The Joint Astronomy Centre on behalf of the Particle Physics and Astronomy Research council of the United Kingdom, the Netherlands Organization for Scientific Research, and the National Research Council of Canada. The VLA is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc. The research of S.R.K. and E.S.P. is supported by the National Science Foundation and NASA.

Correspondence and requests for materials should be addressed to S.R.K. (e-mail: srk@astro.caltech.edu).

An unusual supernova in the error box of the γ -ray burst of 25 April 1998

T. J. Galama¹, P. M. Vreeswijk¹, J. van Paradijs^{1,2}, C. Kouveliotou^{3,4}, T. Augusteijn⁵, H. Bönhardt⁵, J. P. Brewer⁵, V. Doublier⁵, J.-F. Gonzalez⁵, B. Leibundgut⁵, C. Lidman⁵, O. R. Hainaut⁵, F. Patat⁵, J. Heise⁶, J. in 't Zand⁶, K. Hurley⁷, P. J. Groot¹, R. G. Strom^{1,8}, P. A. Mazzali⁹, K. Iwamoto¹⁰, K. Nomoto^{10,11}, H. Umeda^{10,11}, T. Nakamura¹⁰, T. R. Young¹⁰, T. Suzuki^{10,11}, T. Shigeyama^{10,11}, T. Koshut^{3,4}, M. Kippen^{3,4}, C. Robinson^{3,4}, P. de Wildt¹, R. A. M. J. Wijers^{12,13}, N. Tanvir¹², J. Greiner¹⁴, E. Pian¹⁵, E. Palazzi¹⁵, F. Frontera¹⁵, N. Masetti¹⁵, L. Nicastro¹⁶, M. Feroci¹⁷, E. Costa¹⁷, L. Piro¹⁷, B. A. Peterson¹⁸, C. Tinney¹⁹, B. Boyle¹⁹, R. Cannon¹⁹, R. Stathakis¹⁹, E. Sadler²⁰, M. C. Begam²¹ & P. Ianna²¹

The discovery of afterglows associated with γ -ray bursts at X-ray¹, optical² and radio³ wavelengths and the measurement of the redshifts of some of these events^{4,5} has established that γ -ray bursts lie at extreme distances, making them the most powerful photon-emitters known in the Universe. Here we report the discovery of transient optical emission in the error box of the γ -ray burst GRB980425, the light curve of which was very different from that of previous optical afterglows associated with γ -ray bursts. The optical transient is located in a spiral arm of the galaxy ESO184-G82, which has a redshift velocity of only 2,550 km s⁻¹ (ref. 6). Its optical spectrum and location indicate that it is a very luminous supernova⁷, which has been identified as SN1998bw. If this supernova and GRB980425 are indeed associated, the energy radiated in γ -rays is at least four orders of magnitude less than in other γ -ray bursts, although its appearance was otherwise unremarkable: this indicates that very different mechanisms can give rise to γ -ray bursts. But independent of this association, the supernova is itself unusual, exhibiting an unusual light curve at radio wavelengths that requires that the gas emitting the radio photons be expanding relativistically^{8,9}.

GRB980425 was detected¹⁰ on 1998 April 25.91 UT with one of the Wide Field Cameras (WFCs) and the Gamma Ray Burst Monitor on board BeppoSAX, and with the Burst and Transient Source Experiment (BATSE) on board the Compton Gamma Ray Observatory. The BATSE burst profile consisted of a single wide peak. The burst flux rose in ~ 5 s to a maximum flux of $(3.0 \pm 0.3) \times 10^{-7}$ erg cm⁻² s⁻¹ (24–1,820 keV), at which it remained for ~ 5 s; it decayed steadily to the background in ~ 25 s. The burst fluence E_b is $(4.4 \pm 0.4) \times 10^{-6}$ erg cm⁻²; its duration¹¹ T_{90} is 23.3 ± 1.4 s. The burst spectrum is well described by a smoothly broken power law, with a constant break energy $(148 \pm 33$ keV) and high-energy power-law photon index (-3.8 ± 0.7) ; the low-energy power-

law photon index varied from -1.0 ± 0.15 during the rise, to -2.6 ± 0.2 during the decay of the burst. Thus, with respect to its γ -ray properties, GRB980425 was not a remarkable event. In the Beppo SAX WFC no. 2 the burst lasted ~ 30 s, and reached a peak intensity of ~ 3 Crab (2–28 keV)¹⁰. The position derived from the WFC image is right ascension (RA) 19 h 34 min 54 s, declination (dec.) $-52^\circ 49.9'$ (J2000.0), with an error radius of $8'$ which comprises a $3'$ statistical error (99% confidence level) and a $5'$ systematic uncertainty due to incomplete satellite attitude information.

We observed the error box of GRB980425 in R_{MACHO} and B_{MACHO} wavebands¹² with the 50-inch telescope at the Australian National University's (ANU) Mt Stromlo Observatory (MSO) starting April 26.60 UT, and in standard U, B, V, R and I bands with the 30-inch telescope at MSO, the 40-inch telescope at the ANU Siding Spring Observatory, the Anglo-Australian Telescope at the Anglo-Australian Observatory, the 3.5-m New Technology Telescope (NTT), and the 1.5-m Danish and the 0.9-m Dutch telescopes at the European Southern Observatory.

Inspection of NTT images obtained on April 28.4 and May 1.3 UT revealed a point source in the WFC error box, which was not visible in the Digitized Sky Survey. Tying 40-inch V- and R-band images to the Hipparcos Tycho coordinate system, we determined its position at RA 19 h 35 min 03.34 s (± 0.02 s), dec. $= -52^\circ 50' 44.8''$ ($\pm 0.2''$) (J2000.0), $1.6'$ away from the centre of the WFC error box. The source is within the BATSE/Ulysses Interplanetary Network annulus and coincides with the transient radio source in the WFC error box⁸ to within $0.3''$. It is located in an H II region in a spiral arm of the face-on barred spiral galaxy ESO184-G82 at a redshift⁶ of 2,550 km s⁻¹, in the DN1931-529 group of galaxies¹³.

The UBVR light curves of the transient are shown in Fig. 1. These light curves are very different from those of γ -ray burst (GRB) afterglows—which decay as a power law, $F(t) \propto t^{-\alpha}$, with α in the range¹⁴ 1–2—but are quite similar to those of supernovae. This, and the similarity of its spectrum to that of some supernovae (for example SN1994I, Fig. 2) leads us to conclude that the transient is a very luminous supernova of type Ic.

Any estimate of the probability that the supernova and the GRB coincided by chance (with respect to both time and direction) suffers from the problem of *a posteriori* statistics; that is, that the parameters of the problem tend to be set by the observed phenomenon itself. In this case the parameters are the size of the error box, the peak magnitude of the supernova, and the time window within which the events can be considered as possibly related. In our computation we have made generous estimates of these parameters.

The WFC error boxes have 99% confidence level radii varying between $3'$ and $8'$ (ref. 15). We conservatively estimate the angular distance, beyond which a connection can be rejected, at $10'$. We included all supernovae with peak B-band magnitudes $m_B < 16$; that is ~ 2 mag below that of SN1998bw. The time of occurrence of the core collapse and the GRB coincide to within $(+0.7, -2.0)$ days (ref. 16). As a GRB which occurred a few days earlier or later would have been considered at least remarkable, we have taken a time window of 10 days.

With peak absolute magnitudes $M_B - 5 \log h = -18.28$, -16.68 and -15.69 (where h is the Hubble constant in units of 100 km s⁻¹ Mpc⁻¹) for supernovae of types Ia, Ib/c and II, respectively¹⁷, for $m_B < 16$ such supernovae are detectable out to redshifts of 7,180, 3,440 and 2,180 km s⁻¹, respectively. (We note that these limiting values are independent of the assumed value of the Hubble constant.) The Shapley–Ames 'fiducial' sample of 342 galaxies within the Virgo circle¹⁷ has a mean B-band luminosity of $6.7h^{-2} \times 10^9 L_\odot(B)$, and a supernova rate of $3.09h^2[100 \text{ yr } 10^{10} L_\odot(B)]^{-1}$. Using galaxy numbers and heliocentric radial velocities from ref. 18, assuming a mean luminosity, galaxy composition, and supernova rate as in the 'fiducial' sample, and taking relative supernova rates¹⁹ for types II:Ib/c:Ia of 4.0:0.8:1.8, we

¹ Astronomical Institute "Anton Pannekoek", University of Amsterdam & Center for High Energy Astrophysics, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands; ² Physics Department, University of Alabama in Huntsville, Huntsville, Alabama 35899, USA; ³ Universities Space Research Association; ⁴ NASA Marshall Space Flight Center, ES-84, Huntsville, Alabama 35812, USA; ⁵ ESO, Casilla 19001, Santiago 19, Chile; ⁶ SRON Laboratory for Space Research, Sorbonnelaan 2, 3584 CA Utrecht, The Netherlands; ⁷ Space Sciences Laboratory, Berkeley, California 94720-7450, USA; ⁸ Netherlands Foundation for Research in Astronomy, Postbus 2, 7990 AA Dwingeloo, The Netherlands; ⁹ Osservatorio Astronomico di Trieste, Via G.B. Tiepolo 11, I-34131 Trieste, Italy; ¹⁰ Department of Astronomy, School of Science, University of Tokyo, Tokyo 111, Japan; ¹¹ Research Center for the Early Universe, School of Science, University of Tokyo, Tokyo 111, Japan; ¹² Institute of Astronomy, Madingley Road, Cambridge CB3 0HA, UK; ¹³ Department of Physics and Astronomy, SUNY, Stony Brook, New York 11794-3800, USA; ¹⁴ Astrophysikalisches Institut, Potsdam, Germany; ¹⁵ Istituto Tecnologie e Studio Radiazioni Extraterrestri, CNR, Bologna, Italy; ¹⁶ Istituto di Fisica Cosmica e Applicazioni all'Informatica, CNR, Via U. La Malfa 153, I-90146 Palermo, Italy; ¹⁷ Istituto di Astrofisica Spaziale, CNR, Roma, Italy; ¹⁸ Mt Stromlo and Siding Spring Observatories, The Australian National University, Weston Creek, ACT 2611, Australia; ¹⁹ Anglo-Australian Observatory, PO Box 296, Epping, NSW 2121, Australia; ²⁰ School of Physics A29, University of Sydney, NSW 2006, Australia; ²¹ Department of Astronomy, PO Box 3818, University of Virginia, Charlottesville, Virginia 22903, USA.

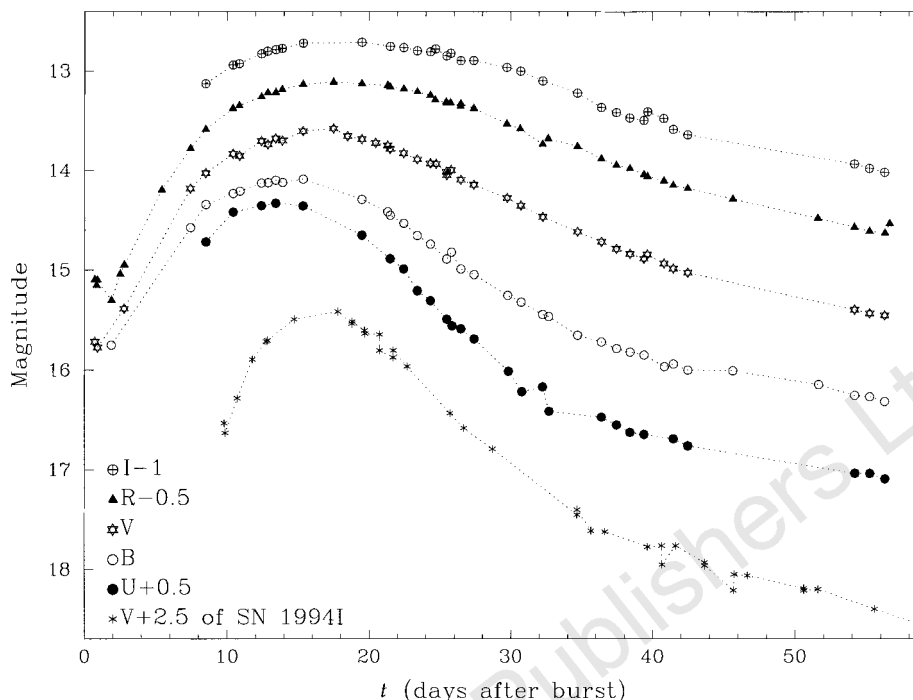


Figure 1 UBVR light curves of SN1998bw, corrected for galactic foreground extinction, $A_V = 0.20$, as inferred from a combination of COBE/DIRBE and IRAS/ISSA maps²⁵. Time is in days since April 25.91 UT (a full log of the observations and the photometry can be found at <http://www.astro.uva.nl/titus>). We determined a photometric (U, B, V, R and I) calibration for a number of reference stars using NTT (May 4.4 UT) and 1.5-m Danish telescope (May 8.3 UT) observations of the Landolt²⁶ fields Mark A and SA110 (stars 496–507) (magnitudes of the reference stars can be found at <http://www.astro.uva.nl/titus>). We corrected for atmospheric extinction and, for U and B, also for a first-order colour term. By comparison of these two calibration nights we estimate an error of the absolute calibration of 0.10 mag in U and 0.05 mag in B, V, R and I. The R_{MACHO} and B_{MACHO} observations have been

transformed using ref. 12. We consider a conservative minimum error of 0.03 mag realistic for the differential U, B, V, R and I light curves to account for the effect of seeing on the contribution of the underlying galaxy (<0.01 mag for each band) and the different instruments used. The longer the wavelength is, the later the maximum light occurs (see Table 1). The R-band light curve shows an initial 'plateau', then it rises at a rate of 0.25 mag d^{-1} to maximum light on May 12. Lack of early data prevents us from establishing the existence of the plateau in the U, B, V and I light curves. Starting early June 1998 the light curves decay exponentially with $\sim 0.025 \text{ mag d}^{-1}$. For comparison the V-band light curve of the type Ic SN1994I is shown²⁷. As discussed in Ref. 16, the width of the light-curve peak depends on the total ejected mass and the explosion energy.

find a total rate of supernovae (with $m_B < 16$ at the peak) of 80 per year. This value includes a correction for absorption¹⁷ within the host galaxy disk. This number should perhaps be increased by a modest factor to account for incompleteness of the radial-velocity distribution¹⁸; we have adopted a final supernova rate ($m_B < 16$) of 120 per year.

With the above parameters, we estimate the probability of catching a supernova in one of the 13 WFC GRB error boxes to be 9×10^{-5} . In our probability estimate we have included all supernovae with peak magnitudes two magnitudes below that of SN1998bw, and we have ignored the fact that SN1998bw is of a rare type. We therefore believe our estimate is conservative. As a result, the notion that GRB980425 and SN1998bw are physically related becomes difficult to reject purely on the basis of the fact that afterglows observed so far from GRBs are very different from supernovae.

The WFC error box contains two X-ray sources^{20,21}, neither of which coincides with SN1998bw. One, 1SAX J1935.0 – 5248 has a constant (2–10 keV) flux of $\sim 2 \times 10^{-13} \text{ erg cm}^{-2} \text{ s}^{-1}$. The other, 1SAX J1935.3 – 5252, was detected at $(1.6 \pm 0.3) \times 10^{-13} \text{ erg cm}^{-2} \text{ s}^{-1}$ about 1 day after the burst, and decayed to $<1.2 \times 10^{-13} \text{ erg cm}^{-2} \text{ s}^{-1}$ (3σ) in 22 hours; it was not detected 6 days after the burst

($<1.0 \times 10^{-13} \text{ erg cm}^{-2} \text{ s}^{-1}$). This variability is consistent with that of previously observed X-ray afterglows of GRBs, and this object might be a possible counterpart for GRB980425. Comparison of the 50-inch April 26.63 UT and April 28.68 UT images at the locations of the two X-ray sources shows no sources variable by more than 0.2 mag down to $R = 21$. However, several GRBs have not shown optical afterglows either, most notably GRB970111²² and GRB970828²³.

The (2–10) keV detection limit (3σ) for the GRB980425 Narrow-Field Instrument observations was $1.2 \times 10^{-13} \text{ erg s}^{-1} \text{ cm}^{-2}$. Using the ASCA (2–10 keV) source count distributions²⁴ one expects to find an average of 0.6 X-ray sources above this limit in the WFC error box; the probability of finding two or more sources there by chance coincidence is 12%. The case for a relation between this X-ray source and GRB980425 must therefore be considered tentative at best, in particular because variability is not rare among weak ROSAT sources.

Modelling¹⁶ of the optical light curve of SN1998bw shows that it can be produced with the core collapse of a massive progenitor star composed mainly of carbon and oxygen (a C + O star); the time of collapse coincides with that of the GRB to within (+0.7, –2.0) days. In the case of the C + O star core collapse, the kinetic energy was $\sim 10^{52.5} \text{ erg}$. To achieve the observed high luminosity, substantial

Table 1 Times of maximum, and apparent and absolute peak magnitudes of SN1998bw

	U	B	V	R	I
Date 1998 (UT)	May 9.4 $\pm$ 0.2	May 10.3 $\pm$ 0.2	May 12.2 $\pm$ 0.2	May 13.2 $\pm$ 0.2	May 13.8 $\pm$ 0.3
Apparent mag	13.81 $\pm$ 0.10	14.09 $\pm$ 0.05	13.62 $\pm$ 0.05	13.61 $\pm$ 0.05	13.70 $\pm$ 0.05
Absolute mag	–19.16 $\pm$ 0.10	–18.88 $\pm$ 0.05	–19.35 $\pm$ 0.05	–19.36 $\pm$ 0.05	–19.27 $\pm$ 0.05

We used $H_0 = 65 \text{ km s}^{-1} \text{ Mpc}^{-1}$, a redshift $z = 2,550 \text{ km s}^{-1}$ and corrected for galactic foreground extinction, $A_V = 0.20$, as inferred from a combination of COBE/DIRBE and IRAS/ISSA maps²⁵.

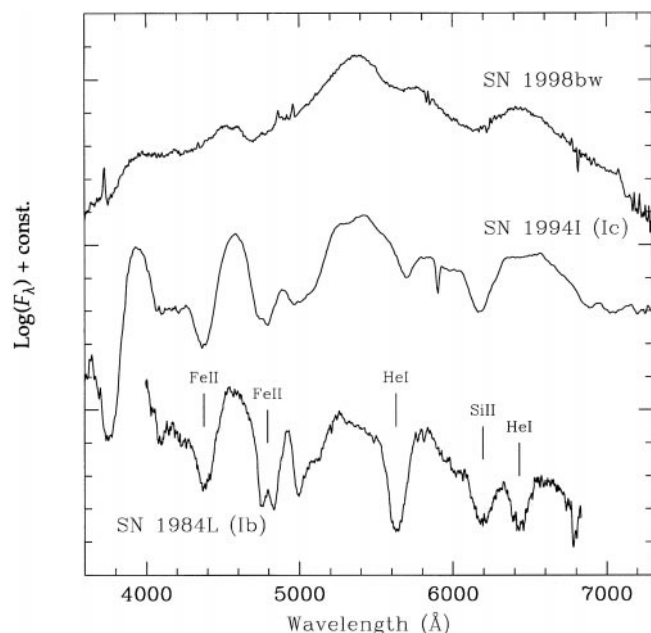


Figure 2 Representative spectra near maximum light of SN1998bw, SN1994I (type Ic; ESO supernova archive, courtesy of M. Turatto), and SN1984L (type Ib)²⁸. Hydrogen lines, characteristic of type II supernovae, and Si II, characteristic of type Ia supernovae, are absent in the spectrum of SN1998bw. The strong He I 5876 line which characterizes type Ib supernovae is very weak in SN1994I and absent in SN1998bw. The overall shape of the spectrum of SN1998bw is similar to that of a type Ic supernova, although the spectral features are less pronounced. The difference is strongest in the 3,500–5,000 Å region, where the Ca II and Fe II lines are much weaker than in SN1994I. In this respect, SN1998bw appears to represent an extreme case in the odd class of type Ic supernovae.

amounts of ^{56}Ni (~ 0.7 solar masses) have to be synthesized in the explosion¹⁶; the large energy and ^{56}Ni mass would be unprecedented for a core-collapse supernova.

If one accepts the possibility that GRB980425 and SN1998bw are associated, one must conclude that GRB980425 is a rare type of GRB, and SN1998bw is a rare type of supernova. The radio properties^{8,9} of SN1998bw show the peculiar nature of this event independent of whether or not it is associated with GRB980425.

The consequence of an association is that the γ -ray peak luminosity of GRB980425 is $L_\gamma = (5.5 \pm 0.7) \times 10^{46} \text{ erg s}^{-1}$ (in the 24–1,820 keV band) and its total γ -ray energy budget is $(8.1 \times 1.0) \times 10^{47} \text{ erg}$. These values are much smaller than those of ‘normal’ GRBs which have peak luminosities of up to $10^{52} \text{ erg s}^{-1}$ and total energies⁵ up to several times 10^{53} erg . This implies that very different mechanisms can produce GRBs which cannot be distinguished on the basis of their γ -ray properties, and that models explaining GRB980425/SN1998bw are unlikely to apply to ‘normal’ GRBs and vice versa. □

Received 11 June; accepted 27 July 1998.

- Costa, E. *et al.* Discovery of an X-ray afterglow associated with the γ -ray burst of 28 February 1997. *Nature* **387**, 783–785 (1997).
- Van Paradijs, J. *et al.* Transient optical emission from the error box of the γ -ray burst of 28 February 1997. *Nature* **386**, 686–688 (1997).
- Frail, D. A., Kulkarni, S. R., Nicastro, L., Feroci, M. & Taylor, G. B. The radio afterglow from the γ -ray burst of 8 May 1997. *Nature* **389**, 261–263 (1997).
- Metzger, M. R. *et al.* Spectral constraints on the redshift of the optical counterpart to the γ -ray burst of 8 May 1997. *Nature* **387**, 878–880 (1997).
- Kulkarni, S. R. *et al.* Identification of a host galaxy at redshift $z = 3.42$ for the γ -ray burst of December 1997. *Nature* **393**, 35–39 (1998).
- Tinney, C., Stathakis, R., Cannon, R. & Galama, T. J. *IAU Circ.* No. 6896 (1998).
- Sadler, E. M., Stathakis, R. A., Boyle, B. J. & Ekers, R. D. *IAU Circ.* No. 6901 (1998).
- Wieringa, M. *et al.* *IAU Circ.* No. 6896 (1998).
- Kulkarni, S. R. *et al.* *IAU Circ.* No. 6903 (1998).
- Soffitta, P. *et al.* *IAU Circ.* No. 6884 (1998).
- Kouveliotou, C. *et al.* Identification of two classes of gamma-ray bursts. *Astrophys. J.* **413**, L101–L104 (1993).

- Bessell, M. S. & Germany, L. M. Calibration of the MACHO photometric system. *Publ. Astron. Soc. Pacif.* (submitted).
- Duis, A. & Newell, B. A catalog of Southern groups and clusters of galaxies. *Astrophys. J. Suppl.* **35**, 209–219 (1977).
- Groot, P. J. *et al.* The rapid decay of the optical emission from GRB980326 and its possible implications. *Astrophys. J.* (in the press).
- Heise, J. *et al.* in *Conf. Proc. 4th Huntsville Symposium on Gamma-Ray Bursts* (eds Meegan, C., Preece, R., Koshut, T.) **428**, 397–403 (Am. Inst. Phys., New York, 1998).
- Iwamoto, K. *et al.* A hypernova model for the supernova associated with the γ -ray burst of 25 April 1998. *Nature* **395**, 672–674 (1998).
- Van den Bergh, S. & Tammann, G. A. Galactic and extragalactic supernova rates. *Annu. Rev. Astron. Astrophys.* **29**, 363–407 (1991).
- Giovannelli, R. & Haynes, M. P. Redshift survey of galaxies. *Annu. Rev. Astron. Astrophys.* **29**, 499–541 (1991).
- Cappellaro, E. *et al.* *Astron. Astrophys.* **322**, 431–441 (1997).
- Pian, E., Frontera, F., Antonelli, L. A. & Piro, L. *GCN Message* No. 69 (1998).
- Pian, E. *et al.* *GCN Message* No. 61 (1998).
- Castro-Tirado, A. *et al.* *IAU Circ.* No. 6598 (1997).
- Groot, P. J. *et al.* A search for optical afterglow from GRB970828. *Astrophys. J.* **493**, L27–L30 (1998).
- Cagnoni, I., Della Ceca, R. & Maccacaro, T. A medium survey of the hard X-ray sky with the ASCA Gas Imaging Spectrometer: the (2–10 keV) number counts relationship. *Astrophys. J.* **493**, 54–61 (1998).
- Schlegel, D. J., Finkbeiner, D. P. & Davis, M. Maps of dust IR emission for use in estimation of reddening and CMBR foregrounds. *Astrophys. J.* (in the press); also as preprint astro-ph/910327 available at (<http://xxx.lanl.gov>) (1998).
- Landolt, A. U. UBVR photometric standard stars in the magnitude range $11.5 < V < 16.0$ around the celestial equator. *Astron. J.* **104**, 340–376 (1992).
- Richmond, M. W. *et al.* UBVR photometry of the type Ic SN 1994I in M51. *Astron. J.* **111**, 327–339 (1996).
- Wheeler, J. C. & Levreault, R. The peculiar type I supernova in NGC 991. *Astrophys. J.* **294**, L17–L20 (1985).

Acknowledgements. This work is based partly on observations made by the MACHO Project with the 50-inch telescope at the ANU’s Mt Stromlo Observatory (ANUSO), by H. Jerjen with the 40-inch telescope at the ANU’s Siding Spring Observatory, and on observations made at the European Southern Observatory, La Silla, Chile. We thank the RAPT Group of amateur astronomers (E. Pozza, A. Brakel, B. Crooke, S. McKeown, G. Wyper, K. Ward, D. Baines, P. Purcell, T. Leach, J. Howard, D. McDowell, M. McDonald, A. Salmon and A. Gurtierrez) for providing images from the 30-inch telescope at ANUSO, and the SuperCOSMOS team for making a scan of an SERC Survey Plate taken with the UKST. J.v.P., C.K., M.K. and K.H. were supported by NASA.

Correspondence and requests for materials should be addressed to T.J.G. (e-mail: titus@astro.uva.nl).

A hypernova model for the supernova associated with the γ -ray burst of 25 April 1998

K. Iwamoto*, P. A. Mazzali†, K. Nomoto‡, H. Umeda*, T. Nakamura*, F. Patat§, I. J. Danziger†, T. R. Young*, T. Suzuki‡, T. Shigeyama‡, T. Augusteijn§, V. Doublie§, J.-F. Gonzalez§, H. Boehnhardt§, J. Brewer§, O. R. Hainaut§, C. Lidman§, B. Leibundgut||, E. Cappellaro*, M. Turatto†, T. J. Galama#, P. M. Vreeswijk#, C. Kouveliotou*, J. van Paradijs††, E. Pian**, E. Palazzi** & F. Frontera**

* Department of Astronomy, ‡ Research Centre for the Early Universe,

School of Science, University of Tokyo, Tokyo 113-0033, Japan

† Osservatorio Astronomico di Trieste, via G.B. Tiepolo 11, I-34131 Trieste, Italy

§ European Southern Observatory, Casilla 19001, Santiago 19, Chile

|| European Southern Observatory, Karl-Schwarzschild-Strasse 2, D-85748,

Garching, Germany

¶ Osservatorio Astronomico di Padova, vicolo dell’Osservatorio 5, I-35122

Padova, Italy

Astronomical Institute ‘Anton Pannekoek’, University of Amsterdam,

and Center for High Energy Astrophysics, Kruislaan 403, 1098 SJ Amsterdam,

The Netherlands

‡ NASA Marshall Space Flight Center, ES-84, Huntsville, Alabama 35812, USA

** Istituto Tecnologie e Studio Radiazioni Extraterrestri, CNR, Bologna, Italy

†† Department of Physics, University of Alabama, Huntsville, Alabama 35899,

USA

The discovery of the unusual supernova SN1998bw, and its possible association with the γ -ray burst GRB 980425^{1–3}, provide new insights into the explosion mechanism of very massive stars and the origin of some classes of γ -ray bursts. Optical spectra indicate that SN1998bw is a type Ic supernova^{3,4}, but its peak luminosity is unusually high compared with typical type Ic supernovae⁵. Here we report our findings that the optical spectra

and the light curve of SN1998bw can be well reproduced by an extremely energetic explosion of a massive star composed mainly of carbon and oxygen (having lost its hydrogen and helium envelopes). The kinetic energy of the ejecta is as large as $(2-5) \times 10^{52}$ erg, more than ten times that of previously observed supernovae. This type of supernova could therefore be termed 'hypernova'. The extremely large energy suggests the existence of a new mechanism of massive star explosion that can also produce the relativistic shocks necessary to generate the observed γ -rays.

SN1998bw is spectroscopically classified as a type Ic supernova, because its optical spectra lack any hydrogen and helium features and the Si II absorption feature is very different from those of type Ia supernovae⁵. Two recent type Ic supernovae, SN1994I^{6,7} and 1997ef⁸, have somewhat similar spectra to that of SN1998bw and their light curves were well reproduced by models of the collapse-induced explosion of C + O stars (Fig. 1). This has led us to construct hydrodynamical models of exploding C + O stars for SN1998bw, assuming that the light is generated by the ^{56}Ni decay as in type Ia supernovae. The model parameters are the stellar mass M_{CO} ,

the explosion energy E_{exp} and the mass of the synthesized ^{56}Ni M_{56} .

Despite their spectral similarity, these three type Ic supernovae have distinctly different brightnesses and light-curve shapes (Fig. 1). This is because the brightness and light-curve shape depend mainly on M_{56} and a pair of $(E_{\text{exp}}, M_{\text{CO}})$, respectively, while the spectral features are sensitive to the chemical composition, which is basically similar among the C + O stars. The peak absolute luminosity $\sim 1.6 \times 10^{43}$ erg s⁻¹ implies that SN1998bw produced $\sim 0.7 M_{\odot}$ of ^{56}Ni , which is much larger than in SN1994I^{6,7} and SN1997ef⁸ (here $M_{\odot}$ is the solar mass).

The above parameters are constrained tightly by comparing the light curves (Fig. 1), synthetic spectra (Fig. 2) and photospheric velocities (Fig. 3) with the observations of SN1998bw³. We find that the optical properties of SN1998bw are best reproduced by a model with $M_{\text{CO}} = 13.8 M_{\odot}$, $E_{\text{exp}} = 3 \times 10^{52}$ erg, and $M_{56} = 0.7 M_{\odot}$ (hereafter designated as CO138). A C + O star of this mass originates from a $\sim 40 M_{\odot}$ main-sequence star. A compact remnant of mass $M_{\text{rem}} = 2.9 M_{\odot}$ must have been left behind for $0.7 M_{\odot}$ ^{56}Ni to be ejected, as required to reproduce the brightness of SN1998bw. This

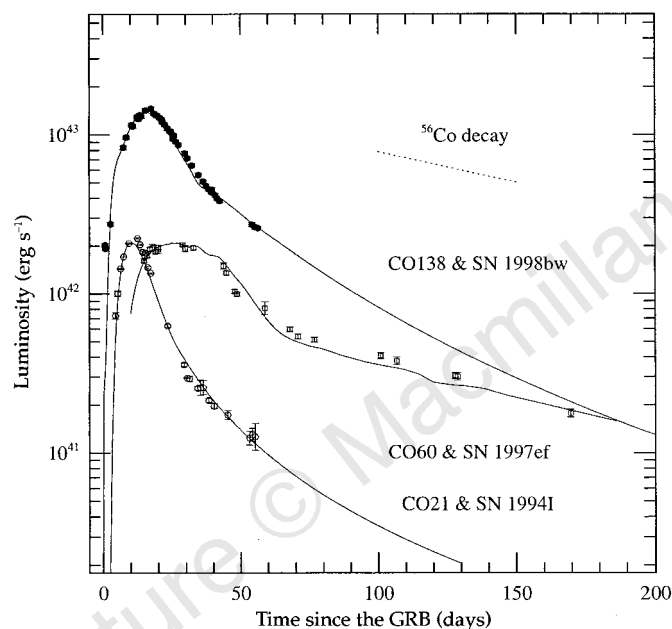


Figure 1 Light curves of three type Ic supernovae, SN1998bw, 1997ef, 1994I and their models. The bolometric light curve of model CO138 ($M_{\text{CO}} = 13.8 M_{\odot}$, $E_{\text{exp}} = 3 \times 10^{52}$ erg, $M_{56} = 0.7 M_{\odot}$) compared with the observations of SN1998bw. The time of the core collapse is set at the detection of GRB980425. The distance to the host galaxy ESO184-G82 is taken to be $\sim 39 \pm 1$ Mpc, as estimated from the redshift $z \approx 0.0085 \pm 0.0002$ and a Hubble constant $65 \text{ km s}^{-1} \text{ Mpc}^{-1}$. The light curves of other type Ic supernovae SN1997ef^{8,16} and SN1994I¹⁷ are also shown, for comparison, together with the corresponding theoretical models, CO60 (ref. 8, $M_{\text{CO}} = 6.0 M_{\odot}$, $E_{\text{exp}} = 10^{51}$ erg, $M_{56} = 0.15 M_{\odot}$) and CO21 (ref. 6, $M_{\text{CO}} = 2.1 M_{\odot}$, $E_{\text{exp}} = 10^{51}$ erg, $M_{56} = 0.07 M_{\odot}$), respectively. We note that E_{exp} and M_{56} of CO138 (SN1998bw) are much larger than in the models for the other two type Ic supernovae. The observed V-band light curves are transformed into the bolometric light curves assuming that the bolometric correction is negligible. The light curves are computed with a radiative transfer code⁹, assuming a detailed balance between photo-ionizations and recombinations and adopting a simplified treatment of line opacity. The explosive nucleosynthesis was calculated using a detailed nuclear reaction network^{18,19} including a total of 211 isotopes up to ^{71}Ge . Our calculation predicts the amount of other radioactive nuclei, $1.4 \times 10^{-3} M_{\odot}$ ^{44}Ti and $1.4 \times 10^{-2} M_{\odot}$ ^{57}Ni , and other stable elements (in $M_{\odot}$) ^{16}O , 7.6; ^{20}Ne , 0.44; ^{23}Na , 1.2×10^{-6} ; ^{24}Mg , 0.46; ^{27}Al , 0.18; ^{20}Si , 0.82; ^{40}Ca , 5.0×10^{-2} ; ^{20}Ne , 0.44.

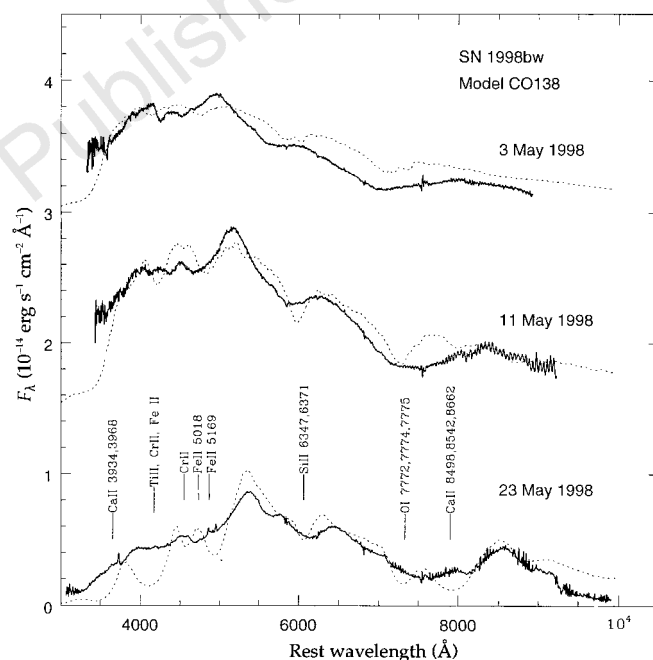


Figure 2 Observed spectra of SN1998bw and synthetic spectra. Three observed spectra (full lines; F.P. *et al.*, manuscript in preparation), where the galaxy background has been subtracted, are compared with the synthetic spectra (dashed lines) computed with a Monte Carlo model¹⁰, improved with the inclusion of photon branching (P.A.M. and L. B. Lucy, manuscript in preparation), using model CO138. The synthetic spectra were computed using the luminosity derived from the light curve and a distance of 39 Mpc, and assuming no reddening. The observed featureless spectra are the result of the blending of many metal lines reaching large velocity and with a large velocity spread. The apparent emission peaks are actually low-opacity regions of the spectrum where photons can escape. The 3 May and the 11 May spectra have been shifted upwards by 3.0×10^{-14} and 1.5×10^{-14} erg s⁻¹ cm⁻² Å⁻¹, respectively. The most important lines are marked on the 23 May spectrum, but they also contribute to the 3 May and 11 May spectra, although with somewhat different ratios. Line blending in the case of SN1998bw is even more severe than it was in the massive type Ic supernova 1997ef⁸, indicating an even larger mass. The massive progenitor model is the only one that gives the correct extent of line blending. Differences in the blue band between the observed spectrum and the synthetic one are probably due to uncertainties in the determination of the abundance and distribution of Fe-group elements in high-velocity parts of the ejecta. The possible presence of O I line absorption in the early spectra complicates any derivation of velocities in the high-velocity wings of the feature conceivably ascribed to Ca II absorption.

mass M_{rem} exceeds the upper mass limit for a stable neutron star, suggesting the formation of a black hole.

To reproduce the light curve of SN1998bw, the time of core collapse should be set to coincide with the detection of GRB980425 to within $+0.7$ – -2 days. The rapid rise of the light curve requires the presence of radioactive ^{56}Ni near the surface, implying that large-scale mixing of material took place, possibly owing to hydrodynamical instabilities. The light-curve shape can be reproduced with different explosion models, because the peak width τ_{LC} , which reflects the timescale of photon diffusion, scales approximately as $\tau_{\text{LC}} \propto \kappa^{1/2} M_{\text{ej}}^{3/4} E_{\text{exp}}^{-1/4}$ (ref. 9), where $M_{\text{ej}} = M_{\text{CO}} - M_{\text{rem}}$ is the mass of the ejected matter, and κ denotes the optical opacity. However, the photospheric velocity scales in a different manner, as $v \propto M_{\text{ej}}^{-1/2} E_{\text{exp}}^{1/2}$, so that both M_{ej} and E_{exp} can be constrained from the spectroscopic data as follows.

Synthetic spectra¹⁰ of various explosion models were compared with the observed spectra of SN1998bw at three epochs: May 3, 11 and 23. The observed featureless spectra are the result of the blending of many metal lines reaching large velocity and with a large velocity spread. Extensive blending can only be achieved with models that have a large mass in high-velocity regions. Therefore, the models that are more massive and have a larger kinetic energy give better fits. For models with $M_{\text{ej}} < 10M_{\odot}$, the photosphere forms at velocities much smaller than those of the observed lines and the lines do not blend as much as in the observed spectra. Figure 2 shows that model CO138 gives consistent fits to the spectra at all three epochs.

Figure 3 shows that the evolution of the photospheric velocity computed from model CO138 (upper solid line) agrees with that obtained from spectral fits (filled circles), and with the observed velocities of the Si II (open circles) and Ca II (squares) lines, within

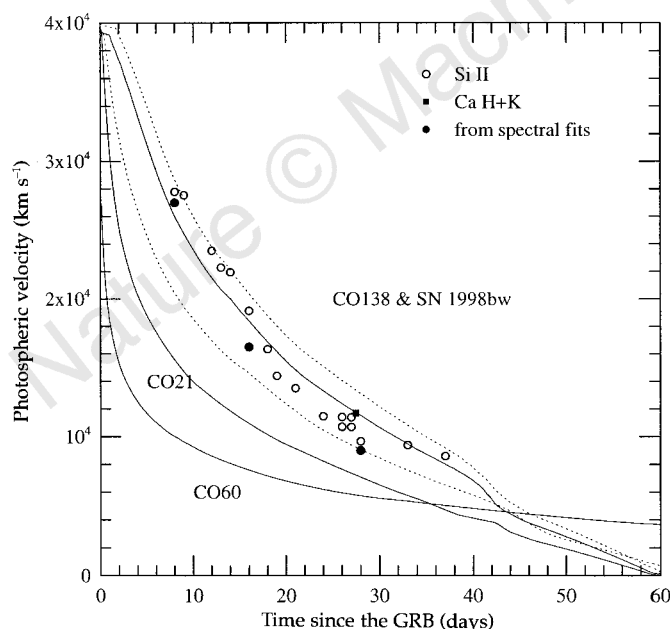


Figure 3 Photospheric velocities of SN1998bw. The evolution of the calculated photospheric velocities of CO138, CO60, and CO21 (solid lines), the photospheric velocity obtained from spectral models (filled circles), the observed velocity of the Si II 634.7, 637.1 nm line measured in the spectra at the absorption core (open circles, F. Patat *et al.*, manuscript in preparation), and that of the Ca II H + K doublet measured in the spectrum of May 23 (square, ref. 4). The observed velocities are in good agreement with the photospheric velocities of CO138, which are much larger than in CO21 and CO60 because of the much larger explosion energies. The upper and lower dotted lines are the velocities of models with $(M_{\text{CO}}, E_{\text{exp}}) = (15M_{\odot}, 5 \times 10^{52} \text{ erg})$ and $(12M_{\odot}, 2 \times 10^{52} \text{ erg})$, respectively. The light curves of these two models also fit SN1998bw well. This indicates the acceptable ranges of $M_{\text{CO}} \approx (12\text{--}15)M_{\odot}$ and $E_{\text{exp}} \approx (2\text{--}5) \times 10^{52} \text{ erg}$.

the uncertainty arising from the light-curve fitting (dotted lines). These velocities are among the highest ever measured in supernovae of any type and thus the smaller-mass C + O star progenitors can be ruled out. By taking into account the uncertainties, we conclude that massive C + O star models with $E_{\text{exp}} \approx (2\text{--}5) \times 10^{52} \text{ erg}$ and $M_{\text{CO}} \approx (12\text{--}15)M_{\odot}$ reproduce well the observed light curve and spectra of SN1998bw.

Here we call the supernova with such an extremely large explosion energy ($>10^{52} \text{ erg}$) a 'hypernova'¹¹. The evolutionary process leading to the hypernova would be as follows. The massive progenitor of initially $\sim 40M_{\odot}$ had a particularly large angular momentum and a strong magnetic field, owing possibly to the spiralling-in of a companion star in a binary system. The collapse of the massive Fe core at the end of the evolution led to the formation of a rapidly rotating black hole. Then the large rotational energy of the black hole was extracted with a strong magnetic field to induce a successful explosion^{11,12}.

The hypernova could induce a γ -ray burst in the following way: at the shock breakout in the energetic explosion, the surface layer is easily accelerated to produce a relativistic shock. When it collides with circumstellar or interstellar matter, non-thermal electrons that are produced at the shock front emit high-energy photons via synchrotron emission. The energy of these photons is given approximately by $160 \text{ keV } (\Gamma/100)^4 n_1^{1/2}$ (ref. 13), where Γ denotes the Lorentz factor of the expanding shell and n_1 is the density of the interstellar matter in cm^{-3} . Thus the event could be observed as a γ -ray burst if Γ becomes as large as ~ 100 . Our preliminary calculations show that spherically symmetric models may not produce large enough energies in γ -rays. However, an axi-symmetric explosion could produce particularly high-speed material by a focused shock wave in a polar direction. The strong radio emission at early phases, which suggests the existence of such a relativistic flow¹⁴, is consistent with the above model. Preliminary spectral polarization measurements show that polarization is small ($\sim 1\%$, but possibly decreasing between 4 and 20 May). Some degree of asymmetry in the envelope morphology is therefore possible, but the precise form depends on the undetermined orientation relative to the line of sight. In the near future, late-time spectra will provide the heavy-element abundances and their velocities in SN1998bw to test our prediction (given in Fig. 1 legend). The late-time decline rate of the light curve is also expected to give further constraints on the model parameters¹⁵. □

Received 11 June; accepted 27 July 1998.

- Soffitta, P., Feroci, M. & Piro, L. *IAU Circ. No. 6884* (1998).
- Lidman, C. *et al. IAU Circ. No. 6895* (1998).
- Galama, T. J. *et al.* An unusual supernova in the error box of the γ -ray burst of 25 April 1998. *Nature* **395**, 670–672 (1998).
- Patat, F. & Piemonte, A. *IAU Circ. No. 6918* (1998).
- Filippenko, A. V. Optical spectra of supernovae. *Annu. Rev. Astron. Astrophys.* **35**, 309–355 (1997).
- Nomoto, K. *et al.* A carbon-oxygen star as progenitor of the type Ic supernova 1994I. *Nature* **371**, 227–229 (1994).
- Iwamoto, K. *et al.* Theoretical light curves for the type Ic supernova SN 1994I. *Astrophys. J.* **437**, L115–L118 (1994).
- Iwamoto, K. *et al.* Light curve modeling of the type Ib/Ic supernova 1997ef. *Astrophys. J.* (submitted).
- Arnett, W. D. Type I supernovae. I. Analytic solutions for the early part of the light curve. *Astron. Astrophys.* **253**, 785–797 (1982).
- Mazzali, P. A. & Lucy, L. B. The application of Monte Carlo methods to the synthesis of early-time supernovae spectra. *Astron. Astrophys.* **279**, 447–456 (1993).
- Paczynski, B. Are gamma-ray bursts in star-forming regions? *Astrophys. J.* **494**, L45–L48 (1998).
- Woosley, S. E. Gamma-ray bursts from stellar mass accretion disks around black holes. *Astrophys. J.* **405**, 273–277 (1993).
- Piran, T. In *Some Unsolved Problems in Astrophysics* (eds Bahcall, J. N. & Ostriker, J. P.) 343–377 (Princeton Univ. Press, 1997).
- Kulkarni, S. R. *et al. IAU Circ. No. 6903* (1998).
- Baron, E., Young, T. R. & Branch, D. *Astrophys. J.* **409**, 417–421 (1993).
- Garnavich, P., Jha, S., Kirchner, R. & Challis, P. *IAU Circ. No. 6798* (1997).
- Richmond, M. W. *et al. UVRI photometry of the Type Ic SN 1994I in M51. Astron. J.* **111**, 327–339 (1996).
- Thielemann, F.-K., Nomoto, K. & Hashimoto, M. Core-collapse supernovae and their ejecta. *Astrophys. J.* **460**, 408–436 (1996).
- Nakamura, T., Umeda, T., Nomoto, K., Thielemann, F.-K. & Burrows, A. Nucleosynthesis in type II supernovae and abundances in metal poor stars. *Astrophys. J.* (submitted).

Acknowledgements. This study is based partly on observations collected at ESO-La Silla.

Correspondence and requests for materials should be addressed to K.N. at the Department of Astronomy (e-mail: nomoto@astron.s.u-tokyo.ac.jp).

Experimental evidence for a two-dimensional quantized Hall insulator

M. Hilke*, D. Shahar*†, S. H. Song*†, D. C. Tsui*, Y. H. Xie‡ & Don Monroe‡

* Department of Electrical Engineering, Princeton University, Princeton, New Jersey, 08544, USA

‡ Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey, 07974, USA

The general theoretical definition of an insulator is a material in which the conductivity vanishes at the absolute zero of temperature. In classical insulators, such as materials with a band gap, vanishing conductivities lead to diverging resistivities. But other insulators can show more complex behaviour, particularly in the presence of a high magnetic field, where different components of the resistivity tensor can display different behaviours: the magnetoresistance diverges as the temperature approaches absolute zero, but the transverse (Hall) resistance remains finite. Such a system is known as a Hall insulator¹. Here we report experimental evidence for a quantized² Hall insulator in a two-dimensional electron system—confined in a semiconductor quantum well. The Hall resistance is quantized in the quantum unit of resistance h/e^2 , where h is Planck's constant and e the electronic charge. At low fields, the sample reverts to being a normal Hall insulator.

Experimentally, the identification of an insulating phase is based on extrapolating the measured magnetoresistance, $\rho_{xx}(T)$, at finite temperature (T) to $T = 0$. This is always an ambiguous process. However, when ρ_{xx} is exponentially increasing as $T \rightarrow 0$, the state is usually considered to be an insulator. Unfortunately, the divergent ρ_{xx} seriously hinders the determination of the Hall resistivity, ρ_{xy} , as even small Hall-contact misalignment will result in a large overriding signal from the diverging ρ_{xx} . It is possible, to a certain degree, to circumvent this difficulty by symmetrizing the measurement. This can be achieved by reversing the magnetic field (B) orientation, as the contribution of ρ_{xx} is symmetric in B as opposed to antisymmetric for ρ_{xy} . The effectiveness of this procedure is demonstrated in the inset of Fig. 1, where we show measurements made in a two-dimensional hole system confined in a 150-Å-thick strained Ge layer sandwiched in between $\text{Si}_{0.4}\text{Ge}_{0.6}$ layers with boron modulation doping. More details of this system are given in ref. 3. The Hall resistances obtained for the two opposite B -field directions are in dotted lines and the average, that is ρ_{xy} , is shown with a solid line. All Hall resistivities discussed here are obtained by this method.

We now turn to discuss our results, where our first task is to identify the different phases. The transition between insulating and quantum Hall phases can be characterized by a critical B -field value, for which ρ_{xx} is T -independent and where the derivative of the T -dependence changes sign on each side of the transition. By plotting ρ_{xx} at two different values of T , we can therefore extract the transition points. In Fig. 1 we have plotted ρ_{xy} together with ρ_{xx} as a function of B . With increasing B , transition points at $B = 2.2$ T and at $B = B_c = 6.06$ T can be identified from the crossing of the two ρ_{xx} curves obtained at different values of T . Between these transitions we have the usual quantum Hall state, which is bordered on both sides by insulators. For clarity, ρ_{xx} is normalized to $\rho_c = \rho_{xx}(B_c)^4$.

Focusing first on the high- B -field region, at $B > B_c$, a striking observation that can be made in Fig. 1 is the large range over which ρ_{xy} remains nearly constant and close to its quantized value of h/e^2 .

Indeed, between 2.7 and 10 T, the deviation of ρ_{xy} from h/e^2 is less than 20%. When doubling the current, the symmetric contribution to ρ_{xy} is further reduced and therefore the deviation is much smaller, that is, less than 5%. This could be due to the good contact alignment as higher currents tend to render the system more homogeneous. In terms of the Landau level filling factor ν , this means that ρ_{xy} remains approximately quantized between $\nu = 1.5$ and $\nu = 0.4$. The transition point is at $\nu_c = 0.75$.

As shown in Fig. 1, we obtained a 5–20% accuracy. To increase this accuracy further but without increasing the current and to analyse the T -dependence, we used a standard low-frequency lock-in technique, with 1 nA currents. In the inset of Fig. 2 we have extracted the T -dependence of ρ_{xy} from the main figure, where ρ_{xx} and ρ_{xy} are plotted as a function of B . The important result here is the clear indication that ρ_{xy} saturates towards the quantized value, within 2%, at low enough temperatures (< 2 K).

We were also interested in studying ρ_{xy} when ρ_{xx} is highly insulating even at low B -field, that is, below the low B -field transition. Figure 2 was therefore obtained by increasing the effective disorder; in other words, we reduced the density of the sample by applying a higher voltage (V_G) on the front gate of our two-dimensional hole system. The low- B -field transition is observed at $B = 2.2$ T in Fig. 1 and at $B = 1.9$ T in Fig. 2, where we have applied a higher V_G . For fields smaller than the value at the low- B -field transition, the system is insulating and ρ_{xy} follows the classical dependence on the magnetic field, namely $\rho_{xy} \approx B/nec$. We observe no temperature dependence of ρ_{xy} for low enough fields. At the same time, $\rho_{xx} \approx 7h/e^2$, at the lowest measured T , indicating that we are in the presence of a strong insulator, more precisely a Hall insulator. To measure only the diffusive part, we limited the value of the maximum phase shift to 2° . Last, we also observed this quantization (within experimental accuracy) when we used the current contacts as Hall probes and the Hall contacts as current

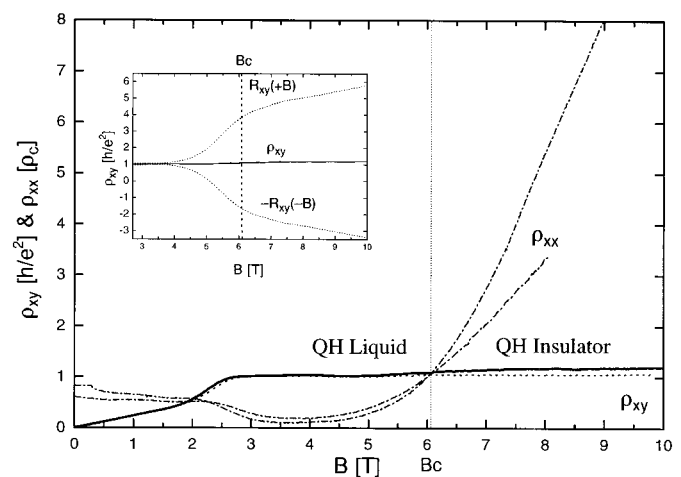


Figure 1 The Hall and diagonal resistivities as a function of B -field. The solid line is the Hall resistivity measured at $T = 300$ mK and a current $I = 200$ nA, whereas the dotted line is for $I = 400$ nA. The dash-dotted lines are ρ_{xx} at $T = 1.2$ K (the uppermost curve) and at $T = 4.2$ K (the lower curve). $V_G = 5.2$ V, $\rho_c = 1.65h/e^2$ and $B_c = 6.06$ T. The inset shows the Hall resistances for $+B$ and $-B$ in dotted lines and ρ_{xy} with a solid line. The experimental error is mainly given by

$$\left| \frac{\Delta R_{xy}}{R_{xy}} \right| \sqrt{1 + \left(\frac{\rho_{xy} - R_{xy}(+B)}{\rho_{xy}} \right)^2}$$

The first term can be estimated from the fluctuations around $\nu = 1$, namely $\sim 4\%$, and the second term is ~ 5 , leading to a total inaccuracy of 20%. With increased current, the first term is reduced to $\sim 3\%$ and the second to ~ 2 , yielding a total inaccuracy of 6%. QH, quantum Hall.

† Present addresses: Department of Condensed Matter Physics, Weizmann Institute, Rehovot 76100, Israel (D.S.); LG Semicon, Choongbuk, 361-480, Korea (S.H.S.).

leads, so a mesoscopic effect can be ruled out. The same results were obtained with a sample from a different wafer (X334) with a very similar structure.

Most previous experiments in the fractional quantum Hall regime^{5–8} suggested a Hall resistance following the classical expression $\rho_{xy} \approx B/nec$. The main difference between these samples and ours is their much higher mobilities, which can account for the different behaviour. In the insulating quantum Hall regime, but with higher mobilities than in our samples, Shahar *et al.*^{9,10} observed that the Hall resistivity remains close to the quantized value around the transition, but at the lowest temperatures studied they observed a small deviation from the quantized value. The origin of this deviation is most probably ρ_{xx} mixing, which becomes too large at very low T and cannot be removed by symmetrizing. In our experiment, however, we managed to minimize this effect, which translates into an even better quantization at low T . In previous experiments, with samples of lower mobilities, ρ_{xy} either had a strong T -dependence¹¹ or a low accuracy¹², and others resolved only the $\nu = 2$ to insulator transition^{11,13,14} together with a small B -field range in the insulating phase, and therefore no generic behaviour was demonstrated. We now present results in which we observed a Hall resistivity equal to the quantized value h/e^2 , within experimental accuracy, as deep as 4 T beyond the transition point $B_C = 6.06$ T.

Although most theories agree on a quantized value of ρ_{xy} at the transition, discrepancies exist for the value inside the insulator. First we point out where the difficulties come from. In standard transport theories, such as in Boltzmann's linear response theory or in the Kubo formalism, one usually calculates conductivities. Using matrix inversion (in two dimensions), one obtains $\rho_{xy} = \sigma_{xx}/(\sigma_{xx}^2 + \sigma_{xy}^2)$. As in an insulator the conductivities vanish, any small correction to the conductivities can alter the value of ρ_{xy} in the limit $\sigma \rightarrow 0$. In the Kubo formulation, the conductivities are expressed as a function of the frequency ω and T . The physically relevant approach, for obtaining the diffusive part of the transport coefficients, is first to take the limit $\omega \rightarrow 0$ and then $T \rightarrow 0$. In the opposite limit, only the reactive part $\sigma_{xx} \sim i\omega$ is obtained^{15,16}, leading to a finite ρ_{xy} , for any type of insulator. In most cases the two limits do not commute.

In the framework of the global phase diagram of Kivelson *et al.*¹, the quantum Hall liquid phases are bordered by an insulator. In the insulating phase, Kivelson *et al.* obtained a finite ρ_{xy} in a random-phase-approximation type of approximation, by evaluating the diffusive part. Concentrating on the transition region, Shimshoni and colleagues^{17,18} considered the duality/particle-hole symmetry, leading to a quantized ρ_{xy} in the critical regime. In this context, the experimental finding of an approximate quantization, throughout the transition, confirms the large region of validity of this duality. When following the analysis of Kivelson *et al.* and considering the insulating phase outside the critical region, an exact quantization would imply the existence of an additional phase transition, that is, from a quantized Hall insulator to Hall insulator behaviour. This would certainly be a very interesting search to pursue in future experiments.

In a different type of approach, a semiclassical network model was considered, yielding a quantized Hall insulator^{2,19}. This network model describes the insulating phase, where quantized Hall liquid states exist in long-range random potential minima. Transport is then obtained via a scattering matrix, parametrizing the tunnelling from one quantized Hall liquid to the other. For the case where phase coherence is destroyed after each scattering event, that is, in the semiclassical limit, ρ_{xy} was shown to remain quantized.

Our main results can be stated as follows: at very low temperatures, and high B -fields, $\rho_{xy} \approx h/e^2$ deep in the insulating phase, hence forming a quantized Hall insulator. At slightly higher temperatures, ρ_{xy} starts to deviate from its quantized value and approaches its classical value proportional to B . At low B -fields, ρ_{xy} follows its classical value even in the strongly insulating case, that is, forming a Hall insulator state. Below, we describe an interesting consequence of this quantization.

Dykhne and Ruzin²⁰ suggested that the conductivities follow a semicircle relation, where each semicircle formed by σ_{xx} as a function of σ_{xy} is centred on $\sigma_{xy} = (n + 1/2)e^2/h$, where n is an integer. We wanted to test this relation for the lowest Landau level, that is, the semicircle centred at $e^2/2h$. Figure 3 presents the result, showing clear evidence for this relation.

This result is in fact not very surprising if we take into account

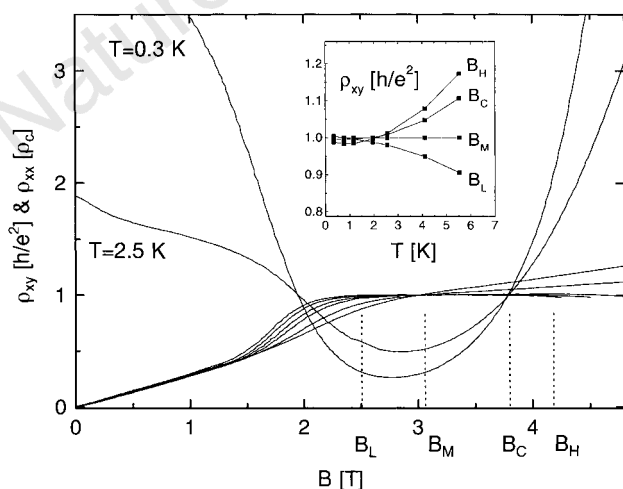


Figure 2 The Hall and diagonal resistivities as a function of B -field for different temperatures. The values of T are 0.3, 1.2, 2, 2.7, 4.2 and 5.5 K. $V_G = 5.4$ V, $\rho_c = 1.73h/e^2$ and $B_C = 3.8$ T. In the inset we have plotted the T -dependence of ρ_{xy} . The constant curve in the inset corresponds to the T -independent ρ_{xy} at $B_M = 3.1$ T, which is close to $\nu = 1$. $B_L = 2.5$ T corresponds to a typical B -field $\nu > 1$, where the Hall resistance decreases with T . $B_C = 3.8$ T corresponds to the critical value defined by the T -independent ρ_{xx} , and $B_H = 4.2$ T to a typical B -field larger than B_C .

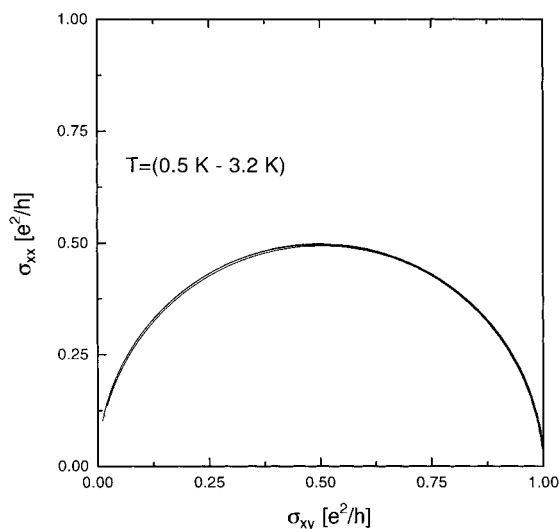


Figure 3 Plot of σ_{xx} as a function of σ_{xy} for different temperatures. The values of T are 0.5, 0.8, 1.2, 1.8 and 3.2 K. $V_G = 5.2$ V and $I = 1$ nA.

the fact that we have a quantized Hall resistance deep inside the insulator. Indeed,

$$\sigma_{xx}^2 + \left(\sigma_{xy} - \frac{e^2}{2h} \right)^2 = \left(\frac{e^2}{2h} \right)^2 + \frac{1 - (e^2/h)\rho_{xy}}{\rho_{xx}^2 + \rho_{yy}^2}$$

simply from matrix inversion. As noted by Shahar *et al.*¹⁰, the semicircle relation follows directly as long as $\rho_{xy} = h/e^2$. In addition, it is completely independent of the value of ρ_{xx} . We emphasize that this relation holds for our entire range of T .

Ando²¹ first calculated numerically the dependence of σ_{xx} on σ_{xy} and found a similar relation to the semicircle function. In the self-consistent Born approximation and scaling theories, a behaviour of this kind is also expected²². A recent numerical study by R. N. Bhatt and S. Subramanian (personal communication), with symmetric and non-symmetric random potentials in a lowest-Landau-level model, confirmed this result. For higher Landau levels, Wei *et al.*²³ and others¹⁰ showed similar dependences. This is, however, the first clear demonstration of this relation for $\nu \leq 1$ and σ_{xx} and σ_{xy} both smaller than $e^2/2h$.

The experiments presented covered a certain range of parameters, in particular in temperature and in disorder. Different behaviours occurring at even lower temperatures or higher mobilities are therefore possible and would be interesting to investigate in future. The existence of these insulators, however, can serve as guidelines for a complete understanding of two-dimensional systems in strong magnetic fields. □

Received 6 April; accepted 14 July 1998.

1. Kivelson, S., Lee, D. H. & Zhang, S. C. Global phase diagram in the quantum Hall effect. *Phys. Rev. B* **46**, 2223–2238 (1992).
2. Shimshoni, E. & Auerbach, A. Quantized Hall insulator: transverse and longitudinal transport. *Phys. Rev. B* **55**, 9817–9823 (1997).
3. Hilke, M. *et al.* Symmetry in the insulator–quantum Hall–insulator transition, observed in a Ge/SiGe quantum well. *Phys. Rev. B* **56**, R15545–R15548 (1997).
4. Shahar, D., Tsui, D. C., Shayegan, M., Bhatt, R. V. & Cunningham, J. E. Universal conductivity at the quantum Hall liquid to insulator transition. *Phys. Rev. Lett.* **74**, 4511–4514 (1995).
5. Willett, R. L. *et al.* Termination of the series of fractional quantum Hall states at small filling factors. *Phys. Rev. B* **38**, 7881–7884 (1988).
6. Goldman, V. J., Shayegan, M. & Tsui, D. C. Evidence for the fractional quantum Hall state at $\nu = 1/7$. *Phys. Rev. Lett.* **61**, 881–884 (1988).
7. Sajoto, T., Li, Y. P., Engel, L. W., Tsui, D. C. & Shayegan, M. Hall resistance of the reentrant insulating phase around the $\nu = 1/5$ fractional quantum Hall liquid. *Phys. Rev. Lett.* **70**, 2321–2324 (1993).
8. Goldman, V. J., Wang, J. K., Su, B. & Shayegan, M. Universality of the Hall effect in magnetic field localized two dimensional electron system. *Phys. Rev. Lett.* **70**, 647–650 (1993).
9. Shahar, D. *et al.* On the nature of the Hall insulator. *Solid State Commun.* **102**, 817–821 (1997).
10. Shahar, D. *et al.* A different view of the quantum Hall plateau-to-plateau transition. *Phys. Rev. Lett.* **79**, 479–482 (1997).
11. Hughes, R. J. F. *et al.* Magnetic-field-induced insulator–quantum Hall–insulator transition in a disordered two-dimensional electron gas. *J. Phys.: Condens. Matter* **6**, 4763–4770 (1994).
12. Kravchenko, S. V., Furneaux, J. E. & Pudalov, V. M. Hall insulator in a two-dimensional electron system in silicon in the extreme quantum limit. *Phys. Rev. B* **49**, 2250–2252 (1994).
13. Jiang, H. W., Johnson, C. E., Wang, K. L. & Hannahs, S. T. Observation of magnetic-field-induced delocalization: transition from Anderson insulator to quantum Hall conductor. *Phys. Rev. Lett.* **71**, 1439–1442 (1990).
14. Lee, C. H., Chang, Y. H., Suen, Y. W. & Lin, H. H. Magnetic-field-induced insulator–quantum Hall conductor–insulator transitions in doped GaAs/AlGaAs quantum wells. *Phys. Rev. B* **56**, 15238–15241 (1997).
15. Viehweger, O. & Efetov, K. B. The Hall coefficient of disordered electronic systems in high magnetic fields. *J. Phys. Condens. Matter* **2**, 7049–7054 (1990).
16. Zhang, S. C., Kivelson, S. & Lee, D. H. Zero temperature Hall coefficient of an insulator. *Phys. Rev. Lett.* **69**, 1252–1253 (1992).
17. Shimshoni, E., Sondhi, S. L. & Shahar, D. Duality near the quantum Hall transitions. *Phys. Rev. B* **55**, 13730–13738 (1997).
18. Shahar, D. *et al.* Evidence for charge flux duality near the quantum Hall liquid-to-insulator transition. *Science* **274**, 589–591 (1996).
19. Pryadko, L. P. & Chalkikian, K. Network of edge states: a random Josephson junction array description. *Phys. Rev. Lett.* **80**, 584–587 (1998).
20. Dykhne, A. & Ruzin, I. Theory of the fractional quantum Hall effect: the two phase model. *Phys. Rev. B* **50**, 2369–2379 (1994).
21. Ando, T. Universal scaling relation of the conductivities in quantum Landau levels. *Surface Science* **170**, 243–248 (1986).
22. Pruisken, A. M. M. Dilute instanton gas as the precursor to the integral quantum Hall effect. *Phys. Rev. B* **32**, 2636–2639 (1985).
23. Wei, H. P., Tsui, D. C. & Pruisken, A. M. M. Localization and scaling in the quantum Hall regime. *Phys. Rev. B* **33**, 1488–1491 (1986).

Acknowledgements. We thank A. Auerbach, L. P. Pryadko, E. Shimshoni and S. L. Sondhi for discussions. This work was supported in part by the National Science Foundation, and M.H. was supported by the Swiss National Science Foundation.

Correspondence should be addressed to M.H. (e-mail: Hilke@ee.princeton.edu).

Room-temperature magnetoresistance in an oxide material with an ordered double-perovskite structure

K.-I. Kobayashi*, T. Kimura*, H. Sawada*, K. Terakura* & Y. Tokura*†

* Joint Research Center for Atom Technology (JRCAT), Tsukuba 305-8562, Japan

† Department of Applied Physics, University of Tokyo, Tokyo 113-8656, Japan

Colossal magnetoresistance—a huge decrease in resistance in response to a magnetic field—has recently been observed in manganese oxides with perovskite structure. This effect is attracting considerable interest from both fundamental and practical points of view¹. In the context of using this effect in practical devices, a noteworthy feature of these materials is the high degree of spin polarization of the charge carriers, caused by the half-metallic nature of these materials^{20,21}; this in principle allows spin-dependent carrier scattering processes, and hence the resistance, to be strongly influenced by low magnetic fields. This type of field control has been demonstrated for charge-carrier scattering at tunnelling junctions^{2,3} and at crystal-twin or ceramic grain boundaries^{4,5}, although the operating temperature of such structures is still too low (≤ 150 K) for most applications. Here we report a material— $\text{Sr}_2\text{FeMoO}_6$, an ordered double perovskite⁶—exhibiting intrinsic tunnelling-type magnetoresistance at room temperature. We explain the origin of this behaviour with electronic-structure calculations that indicate the material to be half-metallic. Our results show promise for the development of ordered perovskite magnetoresistive devices that are operable at room temperature.

In the ordered perovskites $\text{A}_2\text{B}'\text{B}''\text{O}_6$ (where A is an alkaline-earth or rare-earth ion)⁶, the transition-metal sites (perovskite B-sites) are occupied alternately by different cations B' and B'', as depicted in Fig. 1 inset. Intervening oxygen bridges every B' and B'' atom pair, thus forming alternating B'O₆ and B''O₆ octahedra. $\text{Sr}_2\text{FeMoO}_6$ as investigated here is one such example, in which Fe³⁺ (spin quantum number $S = 5/2$) and Mo⁵⁺ ($S = 1/2$) ions order alternately on the perovskite B-sites and the respective spins couple antiferromagnetically. $\text{Sr}_2\text{FeMoO}_6$ has long been known as a conducting ferromagnet (or ferrimagnet) with a fairly high magnetic transition temperature (T_c) of 410–450 K (refs 6–9).

In the perovskite manganites, that is, the prevailing class of compounds that show colossal magnetoresistance, the charge-carriers and local spins arise from the same Mn 3d electron states which are respectively incorporated in the crystal-field-split orbital states but mutually coupled by the strong intra-atomic exchange interaction (Hund's-rule coupling). But in the ordered perovskite $\text{Sr}_2\text{FeMoO}_6$ the local spins of $S = 5/2$ are borne by the Fe³⁺ (3d⁵ electron configuration) ions, while the conduction band is partially occupied by the 4d electrons of Mo⁵⁺. In this case, the antiferromagnetic superexchange interaction between the 3d⁵ $S = 5/2$ spins and 4d¹ $S = 1/2$ spins might produce the large ferrimagnetic magnetization below T_c as observed.

To substantiate such an intuitive view, we have calculated the electronic structure of this ordered perovskite by the density functional method. Figure 1 shows the total density of states with majority 'up' and minority 'down' spins as well as the local density of states for the elements. We note the half-metallic nature of the ground state of this compound: the density of states for the down-spin band is present at the Fermi level, whereas the up-spin band forms a gap at the Fermi level. In accord with the above intuitive view, the occupied up-spin band is mainly composed of Fe 3d

electrons hybridized with oxygen 2*p* states (corresponding to the 3*d*⁵ up-spin configuration) and much less of the Mo 4*d* electrons. The nominal Mo *t*_{2g} and *e*_g up-spin bands are above the Fermi level. By contrast, the down-spin band is mainly occupied by oxygen 2*p* states and around the Fermi level by both the Mo 4*d* *t*_{2g} and Fe 3*d* *t*_{2g} electrons, which are strongly hybridized with oxygen 2*p* states. Such a half-metallic nature gives rise to 100% spin-polarized charge carriers in the ground state. (By similar electronic structure calculations, Pickett¹⁰ has predicted the half-metallic state, and the anti-ferromagnetic state in one case, for some other ordered perovskites.) In view of the fairly high *T*_c (410–450 K), the unusually high spin polarization should be realized even around room temperature (perhaps more than 60% judging from the magnetization data, see below), which makes this compound intriguing in the light of possible application to electromagnetic devices.

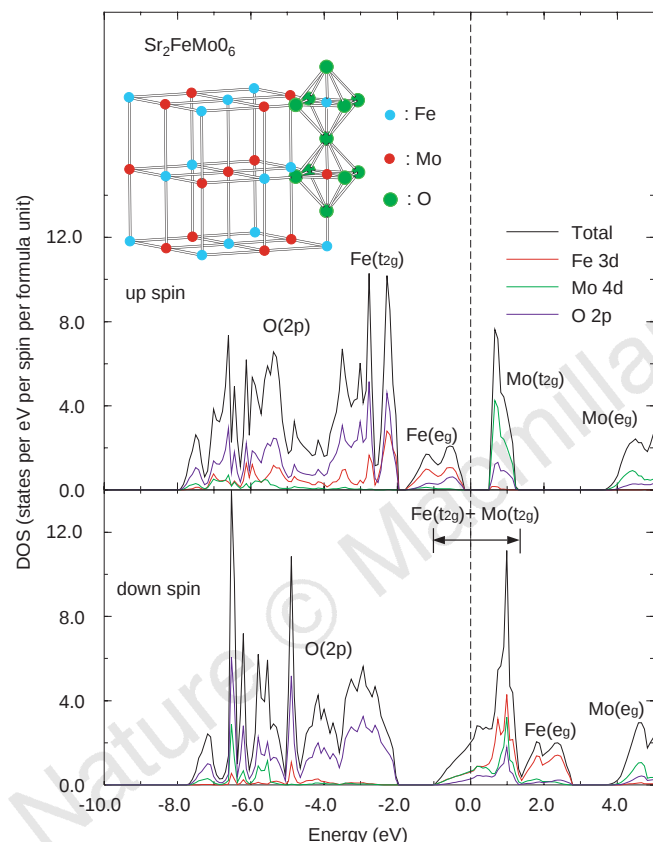


Figure 1 The density of states (DOS) of Sr₂FeMoO₆. The calculation, based on the generalized gradient approximation (GGA)¹⁶ for the exchange and correlation part of the density functional method, was done with the use of the ultrasoft pseudopotential^{17,18} and the plane-wave expansion with a cut-off energy of 30.25 Ry. The lattice parameters are set to be the observed values (see Fig. 2, legend), but the oxygen positions were optimized theoretically. The optimized bond lengths are 2.00 Å (*ab* plane) and 2.01 Å (*c* axis) for Fe–O and 1.94 Å (*ab* plane) and 1.95 Å (*c* axis) for Mo–O. In the up-spin band, the band extending from –8.0 eV to –2.0 eV below the Fermi level is composed mainly of O 2*p* states but the sharp peaks near its upper band edge originate from Fe *t*_{2g} states. The band whose main contribution is Fe *e*_g states is localized between the O 2*p* band and the Fermi level indicated by a broken line at *E* = 0. The sharp narrow band just above the Fermi level is mostly of Mo *t*_{2g} state origin. In the down-spin band, the O 2*p* band is fully occupied and the band from –1 eV to 1.5 eV is contributed by Fe *t*_{2g} and Mo *t*_{2g} states, which is followed by the Fe *e*_g band centred at 2 eV and the Mo *e*_g band starting from 4 eV. The partial density of states of reach of Fe 3*d*, Mo 4*d* and O 2*p* states is also shown, defined as a contribution from the particular state within the spherical region of the corresponding atom whose radius is 2.1 atomic units (1 atomic unit = 0.529 Å) for Fe and Mo, and 1.4 atomic units for O. The calculated magnetic moments for the same ‘muffin-tin’ radii are +3.79 μ_B and –0.29 μ_B, respectively.

The temperature dependence of the resistivity (ρ) for ceramics of Sr₂FeMoO₆ is shown in Fig. 2 at magnetic fields (in T) of 0, 0.2, 0.5, 1, 3 and 7. Figure 2 inset shows the temperature dependence of the magnetization, measured at 1 T. The ferromagnetic (or ferrimagnetic) transition temperature *T*_c of the present sample was determined to be 415 K by the a.c. susceptibility measurement. At zero field, the resistivity shows a very gradual increase with decrease of temperature but never tends to go infinite at the lowest temperature. However, the resistivity value itself is around 0.03 Ω cm, too high for a metallic state. This indicates that the resistance of the sample is dominated by the carrier scattering at the grain boundaries, as widely seen for the polycrystalline ceramic samples⁵. In fact, the resistivity value and its temperature dependence are observed to depend critically on the sample preparation procedure; values ranging from 0.005 to 0.5 Ω cm may be found under various annealing conditions of temperature and atmosphere. Thus the bulk grain of the ceramics should be metallic, but the actual conduction path seems to depend sensitively on the sample form.

As seen in Fig. 2, application of magnetic fields considerably reduces the resistivity over the whole temperature region, but in a gently temperature-dependent manner. In particular, the magnitude of magnetoresistance (MR) is appreciable in a relatively low field region, say below 1 T. Such a low-field-sensitive MR, as well as its gradually-increasing magnitude with decrease of temperature, signals that the spin-dependent scattering is not due to the bulk phenomenon arising from the thermal fluctuation of spins but occurs at the grain boundaries or at the magnetic domain boundaries. We note that the present type of MR has also been observed in loosely sintered polycrystalline samples of the perovskite manganites⁵ and some other half-metallic oxides^{11,12}, as well as in the granular alloys containing some transition metal^{17,18}. In the case

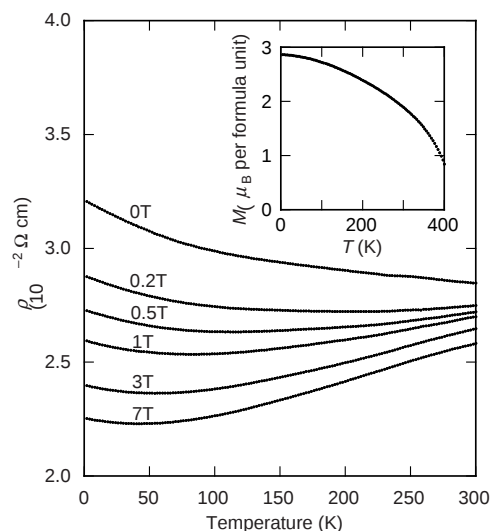


Figure 2 Temperature dependence of the resistivity of the sintered polycrystalline sample of Sr₂FeMoO₆ at zero field and under magnetic fields (in T) of 0.2, 0.5, 1, 3 and 7. The resistivity shows essentially no dependence on the applied field direction. The samples were prepared by solid-state reaction. The mixture of SrCO₃, Fe₂O₃ and MoO₃ was calcined at 900 °C for 3 h in air: the calcined mixture was then finely pulverized and made into pellets followed by sintering at 1,200 °C for 2 h in a stream of 1% H₂/Ar. Analysis of the X-ray powder diffraction pattern of single-phase Sr₂FeMoO₆ in terms of the Rietveld analysis (*R*_{wp} = 9.90%) indicated the 87% order of Fe and Mo on the B-site and a tetragonal lattice structure with *a* = *b* = 5.56850(6) Å and *c* = 7.89722(12) Å. Inset, temperature dependence of the magnetization measured at 1 T. The magnetic transition temperature (*T*_c = 415 K) was determined by the measurement of a.c. susceptibility. The decrease of the resistivity under the magnetic fields, in particular at lower fields (<1 T), is discernible over the whole temperature region below 300 K, and the magnitude of the magnetoresistance gradually increases with decrease of temperature.

of the oxide ceramics^{5,11,12}, the origin of the low-field MR has been interpreted in terms of tunnelling MR at the grain boundaries. In this process, the hopping of the spin-polarized electron between microcrystalline grains (each with their own magnetic domains) is critically affected by the relative angle of the magnetization directions, and hence may be controlled by an external magnetic field via the domain-rotation process. We speculate, on the basis of the observed features, that a similar microscopic process is dominant in the present sample.

To quantify the phenomena observed here, we show the isothermal MR and magnetization curves at 4.2 K and 300 K in Fig. 3a and b, respectively. Corresponding to the results shown in Fig. 2, a large negative MR was observed. Defining the MR magnitude as:

$$MR(T, H) = [\rho(T, H = 0) - \rho(T, H)] / \rho(T, H) \quad (1)$$

We obtain values of MR(4.2 K, 7 T) and MR(300 K, 7 T) as large as 42% and 10%, respectively. We note that a steep low-field MR curve against the field is observed not only at 4.2 K but also at room temperature (300 K). (The same ordinate scale is used for both temperatures in Fig. 3.) By comparison between Fig. 3a and b, we notice that ρ decreases steeply in a low-field region, in good coincidence with the steep magnetization process. Such a close correlation between the MR and the magnetic domain rotation ensures that the observed MR is due to field suppression of the spin-dependent scattering at the grain or domain boundaries, as mentioned above.

In Fig. 4, we show the temperature dependence of the low-field MR, which is exemplified by the values at 0.2 T and normalized at 0 K. If the spin-dependent scattering at the magnetic domain boundaries is totally responsible for the observed MR, the MR magnitude will be scaled by the square of the spin-polarization of the carriers, namely $(M/M_s)^2$, where M_s is the saturation magnetization^{13,14}. The temperature variation of the observed MR magnitude for $\text{Sr}_2\text{FeMoO}_6$ is approximately in accord with that of $(M/M_s)^2$, as shown in the figure. This is rather rare for the MR oxides, including perovskite manganites²⁻⁵, apart from the case of $\text{Ti}_2\text{Mn}_2\text{O}_7$ (but with a much lower T_c , 120 K)¹¹ where the conduc-

tion electron is considered to be primarily in the $\text{Ti } 3d$ state and not in the $\text{Mn } 3d$ state. In Fig. 4, we show for comparison the temperature variation of the low-field intergrain-tunnelling-type MR (normalized by the value at 0 K) as a function of temperature for ceramic samples of $\text{La}_{2/3}\text{Sr}_{1/3}\text{MnO}_3$ ($T_c = 370$ K) and $\text{Ti}_2\text{Mn}_2\text{O}_7$ ($T_c = 120$ K), reported by Hwang *et al.*^{5,11}. We note the advantage of the present ordered double perovskite—with its subsisting tunnelling-type MR behaviour up to room temperature—over these two prototypical compounds; that is, higher T_c and better quality of the intergrain boundary for the spin-polarized conduction electron.

The significantly high T_c (415 K) of $\text{Sr}_2\text{FeMoO}_6$, or the even

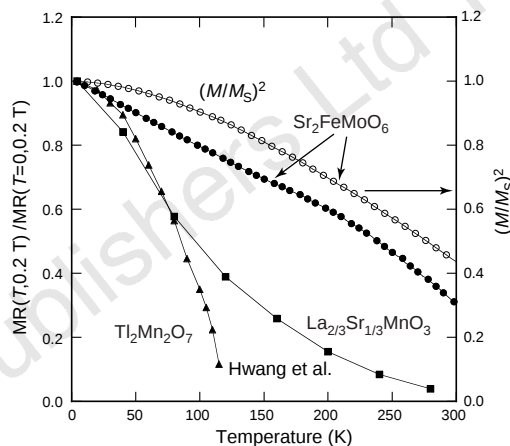


Figure 4 Temperature dependence of the low-field magnetoresistance. Shown is the value of $(MR(T, H) / MR(T = 0, H))$ (see equation (1) in the text) taken tentatively at $\mu_0 H = 0.2$ T and normalized by the value at 0 K, as a function of temperature in comparison with square of the normalized magnetization, $(M/M_s)^2$, where M_s is the saturation magnetization, $3.1 \mu_B$ per formula unit at 4.2 K. The temperature dependence of the intergrain tunnelling-type MR observed for polycrystalline ceramics of perovskite $\text{La}_{2/3}\text{Sr}_{1/3}\text{MnO}_3$ and pyrochlore $\text{Ti}_2\text{Mn}_2\text{O}_7$ by Hwang *et al.*^{5,11} are shown for comparison.

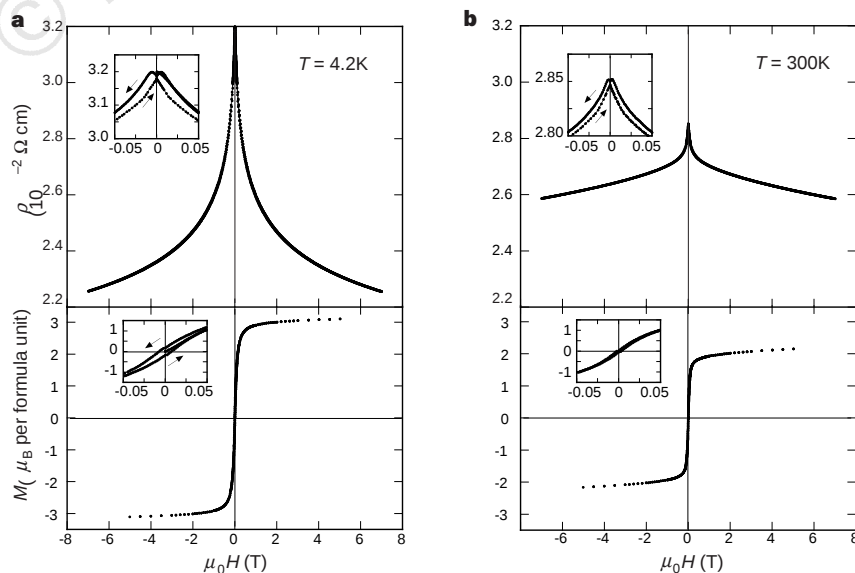


Figure 3 Isothermal magnetoresistance (upper panels) and magnetization M curves (lower panels) for polycrystalline ceramics of $\text{Sr}_2\text{FeMoO}_6$. **a**, At 4.2 K; **b**, at 300 K. The saturation magnetization at 4.2 K is $3 \mu_B$ per formula unit. This confirms the antiferromagnetic coupling between Fe^{3+} ($S = 5/2$) and Mo^{5+} ($S = 1/2$) on the B-sites, as the ferrimagnetic state should give rise to the saturation moment as the difference of the two moments, that is $4 \mu_B$ per formula unit. Small discrepancies between the observed and ideal values might be due to the mis-site-type imperfection of the B-site order, which was partly confirmed by the Rietveld analysis (giving 87% order on the B-site). Following Friedel's arguments for the Ni

alloy (whose d band has a half-metallic feature)⁹, reduction of the magnetic moment caused by one mis-site imperfection is estimated to be $-10 \mu_B$. The magnetization at 300 K nearly saturates around $2.2 \mu_B$ per formula unit, indicating still high spin polarization above 60%, provided that the ground state of this compound shows full spin polarization. The MR magnitude at 7 T is as large as 42% and 10%, at 4.2 K and 300 K, respectively. The insets show a magnification of the field-hysteretic magnetoresistance (upper panels) corresponding to the magnetic hysteresis (lower panels) in a low-field region. The units on the x and y axes are the same as in the main figures.

higher T_c (up to >500 K) of other ordered perovskite compounds^{6,15}, may provide intriguing opportunities to explore magnetoelectronic materials showing half-metallic characteristics and working at room temperature. □

Received 18 May; accepted 27 July 1998.

- Hundley, M. F., Nickel, J. H., Ramesh, R. & Tokura, Y. (eds) *Science and Technology of Magnetic Oxides Mater. Res. Soc. Proc.* **494**, (1998).
- Sun, J. Z. *et al.* Observation of large low-field magnetoresistance in trilayer perpendicular transport devices made using doped manganate perovskites. *Appl. Phys. Lett.* **69**, 3266–3268 (1996).
- Yu, L. *et al.* Large magnetotunneling effect at low magnetic fields in micrometer-scale epitaxial $\text{La}_{0.67}\text{Sr}_{0.33}\text{MnO}_3$ tunnel junctions. *Phys. Rev. B* **54**, R8357–R8360 (1996).
- Mathur, N. D. *et al.* Large low-field magnetoresistance in $\text{La}_{0.7}\text{Ca}_{0.3}\text{MnO}_3$ induced by artificial grain boundaries. *Nature* **387**, 266–268 (1997).
- Hwang, H. Y., Cheong, S.-W., Ong, N. P. & Batlogg, B. Spin polarized tunneling in $\text{La}_{2/3}\text{Sr}_{1/3}\text{MnO}_3$. *Phys. Rev. Lett.* **77**, 2041–2044 (1996).
- Galasso, F. S. *Structure, Properties and Preparation of Perovskite-type Compounds* (Pergamon, London, 1969).
- Patterson, F. K., Moeller, C. W. & Ward, R. Magnetic oxides of molybdenum(V) and tungsten(V) with the ordered perovskite structure. *Inorg. Chem.* **2**, 196–198 (1963).
- Galasso, F. S., Douglas, F. C. & Kasper, R. J. Relationship between magnetic curie points and cell sizes of solutions with the ordered perovskite structure. *J. Chem. Phys.* **44**, 1672–1674 (1966).
- Itoh, M., Ohta, I. & Inaguma, Y. Valency pair and properties of 1:1 ordered perovskite-type compounds Sr_2MMoO_6 (Mn, Fe, Co). *Mater. Sci. Eng. B* **41**, 55–58 (1996).
- Pickett, W. E. Spin-density-functional-based search for half-metallic antiferromagnets. *Phys. Rev. B* **57**, 10613–10619 (1998); Single spin superconductivity. *Phys. Rev. Lett.* **77**, 3185–3188 (1996).
- Hwang, H. Y. & Cheong, S.-W. Low-field magnetoresistance in the pyrochlore $\text{Ti}_2\text{Mn}_2\text{O}_7$. *Nature* **389**, 942–944 (1997).
- Hwang, H. Y. & Cheong, S.-W. Enhanced intergrain tunneling magnetoresistance in half-metallic CrO_2 films. *Science* **278**, 1607–1609 (1997).
- Xiao, J. Q., Jiang, J. S. & Chien, C. L. Giant magnetoresistance in nonmultilayer magnetic systems. *Phys. Rev. Lett.* **68**, 3749–3752 (1992).
- Fujimori, H., Mitani, S. & Ohnuma, S. Tunnel-type GMR in metal-nonmetal granular alloy thin films. *Mater. Sci. Eng. B* **31**, 219–223 (1995).
- Sleight, A. W., Longo, J. & Ward, R. Compounds of osmium and rhenium with the ordered perovskite structure. *Inorg. Chem.* **1**, 245–250 (1962).
- Pedrew, J. P. Atoms, molecules, solids, and surfaces: application of the generalized gradient approximation for exchange and correlation. *Phys. Rev. B* **46**, 6671–6687 (1992).
- Vanderbilt, D. Soft self-consistent pseudopotentials in a generalized eigenvalue formalism. *Phys. Rev. B* **41**, 7892–7895 (1990).
- Lassonen, K., Pasquarello, A., Car, R., Lee, C. & Vanderbilt, D. Car-Parrinello molecular dynamics with Vanderbilt ultrasoft pseudopotentials. *Phys. Rev. B* **47**, 10142–10153 (1993).
- Friedel, J. Metallic alloys. *Nuovo Cimento Suppl.* **2**, 287–311 (1958).
- Okimoto, Y. *et al.* Anomalous variation of optical spectra with spin polarization in double-exchange ferromagnet: $\text{La}_{1-x}\text{Sr}_x\text{MnO}_3$. *Phys. Rev. Lett.* **75**, 109–112 (1995).
- Park, J.-H. *et al.* Direct evidence for a half-metallic ferromagnet. *Nature* **392**, 794–796 (1998).

Acknowledgements. We thank M. Kawasaki, M. Izumi, A. Asamitsu, Y. Tomioka and T. Manako for discussions. This work was partly supported by the New Energy and Industrial Technology Development Organization (NEDO).

Correspondence and requests for materials should be addressed to Y.T. (e-mail: tokura@ap.t.u-tokyo.ac.jp).

Silicon-isotope composition of diatoms as an indicator of past oceanic change

C. L. De La Rocha*†‡, M. A. Brzezinski*†, M. J. DeNiro*§ & A. Shemesh||

* Marine Science Institute, † Department of Ecology, Evolution, and Marine Biology, and § Department of Geological Sciences, University of California, Santa Barbara, California 93106, USA

|| Department of Environmental Sciences and Energy Research, Weizmann Institute of Science, 76100, Rehovot, Israel

Silicon is essential for the growth of diatoms, a group of phytoplankton with opal (amorphous hydrated silica) shells. Diatoms largely control the cycling of silicon in the ocean¹ and, conversely, diatom silica production rates can be limited by the availability of silicic acid². Diatoms are biogeochemically important in that they account for an estimated 75% of the primary production occurring in coastal and nutrient-replete waters¹, rising to more than 90% during ice-edge blooms such as occur in the Ross Sea, off Antarctica³. There are few means by which to reconstruct the

history of diatom productivity and marine silicon cycling, and thus to explore the potential contribution of diatoms to past oceanic biogeochemistry or climate. Indices based on the accumulation of sedimentary opal are often biased by the winnowing and focusing of sediments and by opal dissolution^{4–7}. Normalization of opal accumulation records using particle-reactive natural radionuclides may correct for sediment redistribution artefacts and the dissolution of opal within sediments^{8,9}, but not for opal dissolution before it arrives at the sea floor. Half of the opal produced in the euphotic zone may dissolve before sinking to a depth of 200 m (ref. 1), constituting a potentially large bias to both normalized and uncorrected records of opal accumulation. Here we exploit the potential that variations in the ratio of ^{30}Si to ^{28}Si in sedimentary opal may provide information on past silicon cycling that is unbiased by opal dissolution. Our silicon stable-isotope measurements suggest that the percentage utilization of silicic acid by diatoms in the Southern Ocean during the last glacial period was strongly diminished relative to the present interglacial.

We have demonstrated a fractionation of silicon isotopes of $\sim 1\%$ in $\delta^{30}\text{Si}$ (defined in Fig. 1 legend) during opal formation by marine diatoms that does not vary measurably with temperature (tested from 12–22 °C) or among three species of diatoms investigated⁹. Such discrimination against ^{30}Si during opal formation in a closed system progressively enriches in ^{30}Si the reservoir of dissolved silicic acid utilized for opal formation. Models invoking such Rayleigh process behaviour predict that increased utilization of the silicic acid available for opal formation results in opal with more positive $\delta^{30}\text{Si}$ values, and decreased utilization results in opal with more negative $\delta^{30}\text{Si}$ values⁹. In this manner, $\delta^{30}\text{Si}$ variations in sedimentary opal may reflect changes in the extent of silicic acid use by diatoms in surface waters.

The Rayleigh model requires that opal production be supported by a finite pool of silicic acid, as when diatoms bloom in response to nutrient inputs to a stratified system. In both highly stratified systems with chronically low nutrient levels, and in weakly stratified

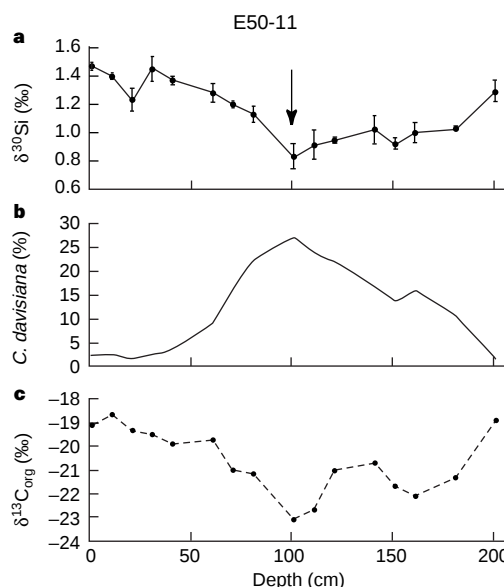


Figure 1 $\delta^{30}\text{Si}$, percentage *C. davisiana* and $\delta^{13}\text{C}_{\text{org}}$ against depth in the southeast Indian core E50-11 (55° 56' S, 104° 57' E, 3,923 m below sea level, m.b.s.l.). Arrows indicate the level of the LGM in the core as determined from *C. davisiana* counts. **a**, Error bars on $\delta^{30}\text{Si}$ represent standard error on 2–3 separate measurements. **b**, Percentage *C. davisiana*²¹. **c**, Values of $\delta^{13}\text{C}_{\text{org}}$, measured from diatom organic matter²². $\delta^{30}\text{Si} = [({}^{30}\text{Si}/{}^{28}\text{Si})_{\text{sample}}/({}^{30}\text{Si}/{}^{28}\text{Si})_{\text{standard}} - 1] \times 10^3$, where the NBS28 standard is used; $\delta^{13}\text{C} = [({}^{13}\text{C}/{}^{12}\text{C})_{\text{sample}}/({}^{13}\text{C}/{}^{12}\text{C})_{\text{standard}} - 1] \times 10^3$, where the PDB standard is used.

‡ Present address: Berkeley Center for Isotope Geochemistry, Department of Geology and Geophysics, University of California, Berkeley, California 94720-4767, USA.

systems where the mixing depth approaches that critical for net phytoplankton growth, a model based on the continuous (relative to utilization) flux of nutrients into the surface layer is more appropriate. In the continuous flux model, the $\delta^{30}\text{Si}$ signature of both opal and ambient silicic acid increases with increasing nutrient utilization until steady state is reached between nutrient inputs and biological removal. Highly stratified open-ocean systems, with nutrient concentrations that are uniform over time and space, are near steady state and the $\delta^{30}\text{Si}$ value of opal there should not vary with silicic acid utilization, but instead should equal that of the silicic acid entering the euphotic zone.

The Southern Ocean, south of the present day Antarctic polar front (APF), is a weakly stratified system that is not at steady state with respect to the physical input of nutrients and their biological removal. Nutrient-rich waters upwell at the Antarctic divergence, and advect northwards to be downwelled at the APF. Silicic acid concentrations in the waters upwelled at this divergence are $>60\text{ }\mu\text{M}$, decreasing to $<10\text{ }\mu\text{M}$ at the APF^{10,11}. The net decrease in silicic acid concentrations along the flow would increase $\delta^{30}\text{Si}$ in opal and silicic acid, thereby following, at least qualitatively, Rayleigh predictions.

$\delta^{30}\text{Si}$ was measured in diatom opal from three Southern Ocean piston cores from locations south of the present day APF: E50-11 and RC11-94 in the Indian sector, and RC13-269 in the Atlantic sector. Diatoms were separated from radiolarians and lithogenic materials, and cleaned of organic and adsorbed matter¹². The resulting 10–30- μm diatom fraction was fluorinated under laser power in a vacuum line¹³, and the $\delta^{30}\text{Si}$ of the resulting SiF_4 gas was measured relative to the standard, NBS28, on a VG Prism II mass spectrometer. Analytical precision on $\delta^{30}\text{Si}$ for replicate samples is $\pm 0.1\text{‰}$.

Profiles of $\delta^{30}\text{Si}$ in all three sediment cores (Figs 1a, 2a, 3a) follow similar trends. Intermediate $\delta^{30}\text{Si}$ values before the onset of full

glacial conditions precede more negative values during the last glacial maximum (LGM). $\delta^{30}\text{Si}$ values then increase up to the Holocene, or present interglacial. The most negative $\delta^{30}\text{Si}$ values in all three cores occur during the LGM (Figs 1a, 2a, 3a). We note that the positions of the LGM in RC11-94 and E50-11 were determined from the relative abundance of the radiolarian *Cycladophora davisiana*^{14,15} and therefore are not as well constrained^{16,17} as the LGM position in RC13-269, which was determined from oxygen-isotope stratigraphy¹⁸.

We propose that variations in silicic acid utilization by diatoms are largely responsible for the downcore variations in $\delta^{30}\text{Si}$. Several lines of evidence support this interpretation. The $\sim 1\text{‰}$ range in $\delta^{30}\text{Si}$ in each core lies within the 1.1‰ range for opal predicted from Rayleigh-process behaviour and the measured fractionation factor⁹. Moreover, carbon isotope variations ($\delta^{13}\text{C}_{\text{org}}$) measured on diatom organic matter from these cores parallel variations in $\delta^{30}\text{Si}$. $\delta^{13}\text{C}_{\text{org}}$ declines with decreasing primary production¹⁹, suggesting that a decline in primary production occurred contemporaneously with diminished utilization of silicic acid (Figs 1c, 2c, 3c). $\delta^{30}\text{Si}$ also visually correlates with the percentage of opal in the cores (Figs 2d, 3d). The percentage of opal, although not in (and of) itself a proxy for opal production, shows the same general patterns as opal accumulation rates in these regions⁴. $\delta^{30}\text{Si}$ thus indicates lower silicic acid utilization during the LGM when opal accumulation and primary productivity were low, and greater utilization with increased productivity and opal accumulation during the Holocene.

Alternative explanations for $\delta^{30}\text{Si}$ variations are species and temperature effects on fractionation, diagenetic alteration of opal $\delta^{30}\text{Si}$, and variations in the $\delta^{30}\text{Si}$ of silicic acid supplied to the euphotic zone. Preliminary results⁹ suggest a lack of species and temperature effects on diatom $\delta^{30}\text{Si}$, although experiments at low temperature with diatoms from the Antarctic remain to be done. We

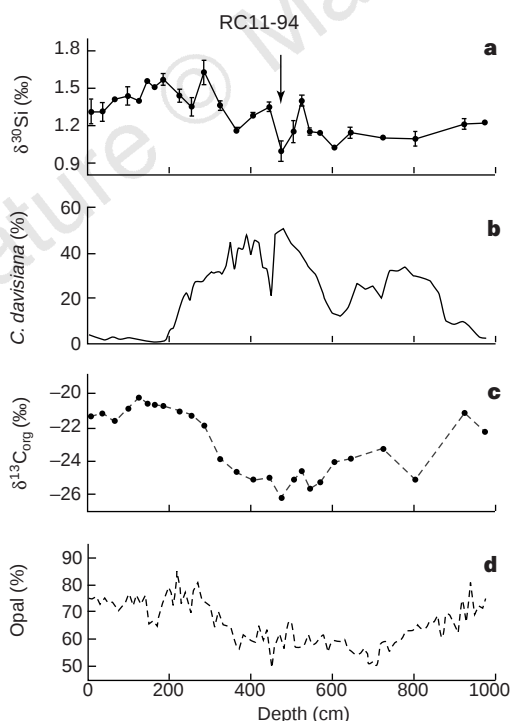


Figure 2 $\delta^{30}\text{Si}$ and other proxies against depth in Indian Ocean core RC11-94 ($54^{\circ}48'\text{S}$, $53^{\circ}03'\text{E}$, 4,303 m.b.s.l.). **a**, $\delta^{30}\text{Si}$; arrow and error bars are as in Fig. 1a. **b**, Percentage *C. davisiana*²¹. **c**, $\delta^{13}\text{C}$ (ref. 31). **d**, Percentage opal (P. N. Froelich, unpublished data), determined through extraction into a sodium carbonate solution³².

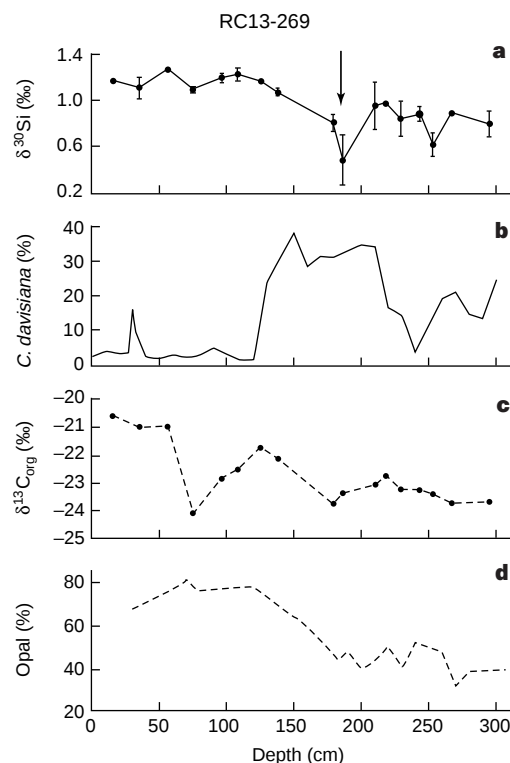


Figure 3 $\delta^{30}\text{Si}$ and other proxies against depth in South Atlantic sediment core RC13-269 ($52^{\circ}38'\text{S}$, $00^{\circ}08'\text{W}$, 2,591 m.b.s.l.). **a**, $\delta^{30}\text{Si}$; arrow and error bars are as in Fig. 1a; arrow falls within the LGM as determined by Charles and Fairbanks¹⁸. **b**, Percentage *C. davisiana*²¹. **c**, $\delta^{13}\text{C}$ against depth (A.S., unpublished data). **d**, Percentage opal (P. N. Froelich, unpublished data), determined through extraction into a sodium carbonate solution³².

also do not believe that post-depositional alteration of $\delta^{30}\text{Si}$ has occurred in these cores. Previous workers concluded that the oxygen-isotope signal in opal from Southern Ocean cores (including RC13-269 and RC11-94) is free from diagenetic artefacts over the past 430 kyr (refs 20, 21). Silicon is less mobile than oxygen in opal, and $\delta^{30}\text{Si}$ should be better preserved than $\delta^{18}\text{O}$. We have also observed that dried diatoms, cleaned of organic matter and resuspended in filtered sea water, retain their $\delta^{30}\text{Si}$ signature even after dissolution of 26% of the opal in their frustules.

Changes in the $\delta^{30}\text{Si}$ of silicic acid supplied to the euphotic zone remain a possible contributor to $\delta^{30}\text{Si}$ variations in sedimentary opal. Such changes might encompass the whole ocean, as suggested for Ge/Si ratios²², or might occur locally with changes in ocean circulation. Whole-ocean changes in $\delta^{30}\text{Si}$ due to variations in terrestrial sources of silicon to the ocean or levels of opal production are unlikely. The range of $\delta^{30}\text{Si}$ in terrestrial igneous materials and clay minerals, the ultimate source of 80% of the silicic acid delivered into the oceans²³, is too small (<1.5‰, ignoring three outlying measurements²⁴) to affect ocean silicic acid $\delta^{30}\text{Si}$ by much. But the sedimentation of opal produced at 90% utilization of silicic acid, compared to the sedimentation of opal produced at 10% utilization, leads to a whole-ocean silicic acid $\delta^{30}\text{Si}$ increase of 0.1‰ per 5,000 years. Silicic acid utilization over the entire ocean is unlikely to change so drastically over glacial cycles, making 0.1‰ per 5,000 years an overestimate.

We may have evidence for variation in the $\delta^{30}\text{Si}$ of silicic acid between ocean basins. There is a 0.3‰ offset between RC13-269 and the Indian sector cores (Figs 1a, 2a, 3a) that exists during both glacial and interglacial times. The magnitude of that offset is similar to the 0.4‰ range (1.1–1.5‰) in the $\delta^{30}\text{Si}$ of silicic acid observed among samples we have obtained from locations as varied as the central North Pacific, a coastal upwelling zone, the North Atlantic, the Southern Ocean, and the subantarctic Atlantic (C.L.D.L.R., unpublished data). Regional differences in the $\delta^{30}\text{Si}$ of the silicic acid supplied to surface waters impede the reconstruction of silicic acid utilization as model results vary directly with the value of this parameter. However, congruency in downcore $\delta^{30}\text{Si}$ variations in all cores suggests that these regional differences persist over glacial cycles, preserving the validity of $\delta^{30}\text{Si}$ variations within each core as a qualitative indicator of silicic acid utilization.

The minimum in silicic acid utilization at the LGM indicated by $\delta^{30}\text{Si}$ (Figs 1a, 2a, 3a) is consistent with the conclusions of several studies of Southern Ocean palaeoproductivity. These studies, based on production proxies such as opal accumulation^{4,5} and diatom $\delta^{13}\text{C}_{\text{org}}$ (ref. 19), and on particle flux proxies such as $^{231}\text{Pa}/^{230}\text{Th}$, $^{10}\text{Be}/^{230}\text{Th}$ and authigenic uranium accumulation⁶, concluded that productivity south of the present APF was lower during the last glacial than it is at present. Increased coverage of this area by sea ice during glacials has been proposed^{4,5,15,25,26}, and would explain the concurrent drop in both productivity and silicic acid utilization at the LGM.

Whereas the $\delta^{30}\text{Si}$ proxy of silicic acid utilization appears to be in phase with the production and particle-flux proxies, $\delta^{30}\text{Si}$ and the $\delta^{15}\text{N}$ -based proxy of nitrate utilization show opposite patterns. South of the present APF, $\delta^{15}\text{N}$ indicates increased utilization of nitrate by phytoplankton at the LGM²⁷, whereas $\delta^{30}\text{Si}$ suggests decreased utilization of silicic acid by diatoms. The antiphasing of these proxies may reflect a shift in the phytoplankton community composition during the LGM that favoured non-siliceous phytoplankton over diatoms, although evidence of this remains to be observed in the sedimentary record. Alternatively, increased availability of Fe during the last glacial in the Southern Ocean⁸ may have contributed to the decoupling of the Si and N records. Diatom Si/N utilization ratios are up to three times higher under iron-limitation than under metal-replete conditions^{28,29}. Iron-enrichment during the LGM could induce a depletion of nitrate with respect to silicic acid without a decline in the relative dominance of diatoms.

Variations in $\delta^{30}\text{Si}$ do not support the glacial increase in surface-water stratification south of the present day APF that has been invoked to explain lower atmospheric CO_2 concentrations during the last glacial period²⁷. Increased stratification of glacial surface waters should result in relatively high values of $\delta^{30}\text{Si}$ at the LGM, decoupling the silicon-isotope signal from the opal accumulation record. Such decoupling is not seen in either of the cores for which we have percentage opal data (Figs 2, 3), although this conclusion awaits verification from opal accumulation records normalized to particle-reactive nuclides such as ^{230}Th or ^{231}Pa (refs 6, 8). High opal accumulation rates in Southern Ocean sediments⁴ show that stratification sufficient to satisfy the critical depth criterion for net diatom growth³⁰ occurs during both the LGM and the Holocene. The minimum in $\delta^{30}\text{Si}$ at the LGM suggests that stratification events were not as numerous or as persistent during the LGM, resulting in less utilization of silicic acid supplied to surface waters. Alternatively, like the northward advance of both the APF and sea ice during the LGM^{4,5,15,25,26}, the Antarctic divergence may migrate northwards during glacials, shortening the distance between the site of upwelling and the location of the sediment cores. Such a decrease in the residence time of upwelled waters would allow less time for diatom growth.

The ability to estimate silicon utilization quantitatively from $\delta^{30}\text{Si}$ would significantly increase the utility of this proxy. The change in utilization was calculated using the measured silicon isotope fractionation factor of 0.9989 (ref. 9), a $\delta^{30}\text{Si}$ value of 1.6‰ for silicic acid supplied to the euphotic zone, and by assuming that the sedimentary opal was produced under bloom conditions. A $\delta^{30}\text{Si}$ value of 1.6‰ falls within the range of the few silicic acid $\delta^{30}\text{Si}$ values available from intermediate depths (C.L.D.L.R., unpublished data) that could supply silicic acid to the euphotic zone. The estimated utilization of silicic acid increased from average values of 40% during the LGM to 90% utilization during the early Holocene. This 125% increase may overestimate the change in silicon utilization, as the exact contribution to downcore variations in opal $\delta^{30}\text{Si}$; values of regional and glacial–interglacial variations in silicic acid $\delta^{30}\text{Si}$ is not yet known. Detailed investigations into the distribution of silicon isotopes in the water column should refine our ability to quantify past variations in absolute levels of silicic acid use. Meanwhile, variations in $\delta^{30}\text{Si}$ values in diatom opal provide an improved qualitative index of silicic acid use for palaeoceanographic studies. □

Received 28 October 1997; accepted 28 July 1998.

- Nelson, D. M., Tréguer, P., Brzezinski, M. A., Leynaert, A. & Quéguiner, B. Production and dissolution of biogenic silica in the ocean: Revised global estimates, comparison with regional data and relationship to biogenic sedimentation. *Glob. Biogeochem. Cycles* **9**, 359–372 (1995).
- Brzezinski, M. A. & Nelson, D. M. Chronic substrate limitation of silicic acid uptake rates in the western Sargasso Sea. *Deep-Sea Res. II* **43**, 437–453 (1996).
- Nelson, D. A. & Smith, W. O. Jr Phytoplankton bloom dynamics of the western Ross Sea Ice edge II. Mesoscale cycling of nitrogen and silicon. *Deep-Sea Res. A* **33**, 1389–1412 (1986).
- Charles, C. D., Froelich, P. N., Zibello, M. A., Mortlock, R. A. & Morley, J. J. Biogenic opal in Southern Ocean sediments over the last 450,000 years: Implications for surface water chemistry and circulation. *Paleoceanography* **6**, 697–728 (1991).
- Mortlock, R. A. *et al.* Evidence for lower productivity in the Antarctic Ocean during the last glaciation. *Nature* **351**, 220–223 (1991).
- Kumar, N. *et al.* Increased biological productivity and export production in the glacial Southern Ocean. *Nature* **378**, 675–680 (1985).
- Dymond, J., Collier, R., McManus, J., Honjo, S. & Manganini, S. Can the aluminum and titanium contents of ocean sediments be used to determine the paleoproductivity of the oceans? *Paleoceanography* **12**, 586–593 (1997).
- Kumar, N., Gwiazda, R., Anderson, R. F. & Froelich, P. N. $^{231}\text{Pa}/^{230}\text{Th}$ ratios in sediments as a proxy for past changes in Southern Ocean productivity. *Nature* **362**, 45–48 (1993).
- De La Rocha, C. L., Brzezinski, M. A. & DeNiro, M. J. Fractionation of silicon isotopes by marine diatoms during biogenic silica formation. *Geochim. Cosmochim. Acta* **61**, 5051–5056 (1997).
- Le Jehan, S. & Tréguer, P. Uptake and regeneration of Si:N:P ratios in the Indian Ocean. *Polar Biol.* **2**, 127–136 (1983).
- Van Bennekom, A. J., Berger, W. H., Van Der Gaast, S. J. & De Vries, R. T. P. Primary productivity and the silica cycle in the Southern Ocean (Atlantic sector). *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **67**, 19–30 (1988).
- Shemesh, A., Mortlock, R. A., Smith, R. J. & Froelich, P. N. Determination of Ge/Si in marine siliceous microfossils: Separation, cleaning and dissolution of diatoms and radiolaria. *Mar. Chem.* **25**, 305–323 (1988).
- De La Rocha, C. L., Brzezinski, M. A. & DeNiro, M. J. Purification, recovery, and laser-driven fluorination of silicon from dissolved and particulate silica for the measurement of natural stable isotope abundances. *Analyt. Chem.* **68**, 3746–3750 (1996).

14. Hays, J. D., Lozano, J. A., Shackleton, N. & Irving, G. Reconstruction of the Atlantic and western Indian Ocean sectors of the 18,000 B.P. Antarctic Ocean. *Geol. Soc. Am. Mem.* **145**, 337–372 (1976).
15. Cooke, D. W. & Hays, J. D. in *Antarctic Geoscience* (ed. Craddock, C.) 1017–1025 (Univ. Wisconsin Press, Madison, 1982).
16. Burckle, L., De Maurel, K. & McHugh, C. G. Latest Quaternary paleoceanography of the Atlantic sector of the Southern Ocean. *Antarct. J. US* **21**, 138 (1986).
17. De Maurel, K., Burckle, L. H. & McHugh, C. G. Late Quaternary paleoenvironments in the Atlantic sector of the Southern Ocean. *Geol. Soc. Am. Abstr.* **18**, 583 (1986).
18. Charles, C. D. & Fairbanks, R. G. in *Geological History of the Polar Oceans: Arctic Versus Antarctic* (eds Bleil, U. & Theide, J.) 519–538 (Kluwer Academic, Bremen, 1990).
19. Shemesh, A., Macko, S. E., Charles, C. D. & Rau, G. H. Isotopic evidence for reduced productivity in the glacial Southern Ocean. *Science* **262**, 407–410 (1993).
20. Shemesh, A., Burckle, L. H. & Hays, J. D. Meltwater input to the Southern Ocean during the last glacial maximum. *Science* **266**, 1542–1544 (1994).
21. Shemesh, A., Burckle, L. G. & Hays, J. D. Late Pleistocene oxygen isotope records of biogenic silica from the Atlantic sector of the Southern Ocean. *Paleoceanography* **10**, 179–196 (1995).
22. Froelich, P. N. *et al.* River fluxes of dissolved silica to the ocean were higher during glacials: Ge/Si in diatoms, rivers, and oceans. *Paleoceanography* **7**, 739–767 (1982).
23. Tréguer, P. *et al.* The silica balance in the world ocean: A reestimate. *Science* **268**, 375–379 (1995).
24. Douthett, C. B. The geochemistry of the stable isotopes of silicon. *Geochim. Cosmochim. Acta* **46**, 1449–1458 (1982).
25. Burckle, L. H. & Cirilli, J. Origin of diatom ooze belt in the Southern Ocean: Implications for late Quaternary paleoceanography. *Micropaleontology* **33**, 82–86 (1987).
26. Shemesh, A., Burckle, L. H. & Froelich, P. N. Dissolution and preservation of Antarctic diatoms and the effect on sediment thanatocoenoses. *Quat. Sci.* **31**, 288–308 (1989).
27. Francois, R. *et al.* Contribution of Southern Ocean surface water stratification to low atmospheric CO₂ concentrations during the last glacial period. *Nature* **389**, 929–935 (1997).
28. Hutchins, D. A. & Bruland, K. W. Iron-limited diatom growth and Si:N uptake ratios in a coastal upwelling. *Nature* **393**, 561–564 (1998).
29. Takeda, S. Influence of iron availability on nutrient consumption ratio of diatoms in oceanic waters. *Nature* **393**, 774–777 (1998).
30. Sverdrup, H. U. On conditions for the vernal blooming of phytoplankton. *J. Cons. Perm. Int. Explor. Mer* **18**, 287–295 (1953).
31. Singer, A. J. *Diatom ¹³C Records from Southern Ocean Sediments: Glacial-interglacial Variation in Surface Water Characteristics*. Thesis, Weizmann Institute of Science (1994).
32. Mortlock, R. A. & Froelich, P. N. A simple method for the rapid determination of biogenic opal in pelagic marine sediments. *Deep-Sea Res.* **36**, 1415–1426 (1989).

Acknowledgements. We thank D. J. DeMaster and P. N. Froelich for reviews; P. N. Froelich for the per cent opal data; D. W. Lea, H. Berg, L. K. Crowder, W. Golden, J. Kennett, R. L. Ripperdan and K. Shea for advice and support; and the curator of the Lamont-Doherty Geological Observatory for samples. This work was supported by the NSF.

Correspondence and requests for materials should be addressed to C.L.D. (e-mail: delaroch@uclink4.berkeley.edu).

Formation of atmospheric particles from organic acids produced by forests

Ilias G. Kavouras, Nikolaos Mihalopoulos & Euripides G. Stephanou

Environmental Chemical Processes Laboratory, Department of Chemistry, University of Crete, 71409 Heraclion, Greece

Aerosol formation in the atmosphere is an important process to understand, in that such particles may act as the cloud condensation nuclei responsible for the 'cloud-climate' effect¹, and could locally be hazardous to health. The number-concentration of total atmospheric aerosols and cloud condensation nuclei is largely contributed by organic aerosols¹. Much of the organic aerosol is formed from atmospheric gas-to-particle conversion, and the common and widespread non-methane hydrocarbons emitted by vegetation have been investigated as possible precursors². But strong evidence for a quantitative link between biogenic hydrocarbon emission and organic aerosol formation has so far been lacking. Here we present measurements of gaseous and particulate atmospheric species from a forested area to show that some hydrocarbons (for example, terpenes) emitted by vegetation are photo-oxidized to organic acids (for example, pinonic acids), which condense to form organic aerosols. Thus the forests, through their production of large quantities of organic aerosols, could be of considerable significance both for climate (through cloud-condensation-nuclei formation) and for heterogeneous atmospheric chemical processes.

Sampling was performed from 10 to 16 August 1996, in an *Eucalyptus globulus* forest located in Tabua (Portugal), situated 100 km inland from the Atlantic coast. Simultaneous measurements of the concentrations of O₃, NO_x, non-methane hydrocarbons, isoprene, terpenes, Aitken nuclei (fine aerosols) and their size distribution in the 3–200 nm range, aerosols and meteorological parameters were performed on a 20-m-high sampling tower, which was situated in the middle of the forest³.

Figure 1 presents the variation of the concentration of Aitken nuclei and their mean diameter during the sampling period; this figure also shows the three intensive sampling periods (I, II and III). The mean diameter of Aitken nuclei and their concentrations follow completely different patterns. For example, the Aitken nuclei concentration remains constant during the night and increases during daytime by reaching its maximum in the afternoon, whereas their mean diameter shows the opposite trend by reaching the lowest values in the afternoon. Taking into account the fact that at the time of the concentration maxima of the Aitken nuclei, the mean particle diameter is in the range 20–40 nm (particles with a lifetime of few minutes), the above observation indicates a local particle source, rather than a long-range transport. New particle formation in a forested area has also been reported⁴, although no chemical analysis of the particulate phase was performed.

Air trajectory analysis showed the existence of two well-distinct situations: the first from the beginning of the experiment (on 13 August) up to 19 August and the second from 19 August until the end of the experiment on 26 August. During the first period the air masses had mainly local influence. They stayed for more than 2–3 days above the Iberian peninsula and reached the sampling station from northeast–east directions. On 19 August, the passage of a front associated with a rain event changed the origin of the air masses: they now originated from the Atlantic Ocean. In Fig. 2 are shown the variations of Aitken nuclei concentrations and their mean diameter (Fig. 2a–c) during the three 48-hour intensive measurement periods (I, II and III; Fig. 1), and for the same periods the corresponding concentration variations (Fig. 2d–f) of the non-seasalt (anthropogenic) sulphate (n.s.s.-SO₄²⁻)⁵. The change of the air-mass origin was clearly reflected in the variation of the concentration of the anthropogenic n.s.s.-SO₄²⁻ mean concentration progressively decreased from 7 µg m⁻³ during the first intensive sampling, to 4.7 µg m⁻³ during the second intensive sampling (associated with the passage of the frontal system) and to 0.5 µg m⁻³ during the third intensive sampling. Despite the clear change in the origin of air masses and the n.s.s.-SO₄²⁻ concentration (considered as one of the main contributors of aerosol formation¹), no major change had occurred in the pattern of the Aitken nuclei concentration (Fig. 2a–c). This fact is consonant with the hypothesis presented above; that is, that new particle formation is due to local atmospheric chemistry phenomena, rather than to a long-range transport. The rapid particle formation observed immediately after the end of rain events (occurred on 19 and 23 August, Fig. 1) is also in agreement with the above hypothesis.

The carbonyl compound fraction of an aerosol sample extract contained pinonaldehyde (2,2-dimethyl-3-acetyl-cyclobutyl-ethanal) and nopinone (6,6-dimethyl-bicyclo[3.1.1]heptan-2-one). Pinonaldehyde and nopinone were identified by comparison of their mass spectra with those of authentic standards (nopinone) and reference mass spectra (pinonaldehyde)^{6,7}. These two compounds were more abundant (Table 1) than the plant leaf wax *n*-alkanals found also in aerosols⁸. Pinonaldehyde and nopinone have been described as photo-oxidation reaction products of α -pinene^{6,9–12} and β -pinene^{6,9,13}, respectively. Pinonaldehyde has been reported to be an important product from the reactions of α -pinene with NO₃ radical (ref. 14), the OH radical and O₃ (refs 6, 9–13). By far the most abundant compound in the acidic fraction of the aerosol organic extract was *cis*-pinonic acid (*cis*-2,2-dimethyl-3-acetyl-cyclobutyl-ethanoic acid). In addition, we identified

trans-pinonic acid (*trans*-2,2-dimethyl-3-acetyl-cyclobutyl-ethanoic acid). The above compounds were also identified by comparing their mass spectra with those of authentic standards and with reference spectra^{6,7}. These two carboxylic acids are formed primarily by the reaction of ozone with α -pinene^{10,12,13,15}. Pinonic acid has been found to be one of the main photo-oxidation reaction products of α -pinene^{6,7,10,13} in chamber experiments under simulated atmospheric conditions, but has been previously identified only once as a component of forest aerosol⁷.

Interconversion of the pinonic acid isomers in the presence of ultraviolet radiation is supported by the fact that during daylight the mean concentration ratio (for periods I and II) of *trans*- to *cis*-pinonic acid, deduced from concentrations in Table 1, was 0.43 ± 0.04 (hours 8–14) and 0.47 ± 0.09 (hours 14–20) while during the night this ratio decreased to 0.14 and 0.29 (hours 20–32). In the acidic fraction, these two photo-oxidation products were more abundant (Table 1) than the plant leaf wax *n*-alkanoic acids. In addition, the analysis of the impactor filters indicated that *cis*- and *trans*-pinonic acids had accumulated only on particles with diameter <500 nm. Particles with diameter >500 nm did not

contain any detectable quantity of the above acids; they contained the leaf wax *n*-alkanoic acids. Although in most published investigations (using chamber experiments) pinonic acid was considered to be a component of the secondary organic aerosols formed through photo-oxidation of α -pinene^{6,10}, its unambiguous presence has been clearly demonstrated in chamber experiments only very recently¹³. Our *in situ* results clearly demonstrate for the first time, to our knowledge, that under real atmospheric conditions the most important compounds in the fine (particles with diameter <500 nm) organic aerosol phase are the two carboxylic acids (*cis*- and *trans*-pinonic acid; Table 1). These two acids represent 18.4–40.5% (Table 1) of the fine particle mass during daytime, when Aitken nuclei reach their maximal concentration.

The variation of concentrations of gaseous formic acid and particulate *cis*- and *trans*-pinonic acid, pinonaldehyde and nopinone is shown in Fig. 3a–d. Photo-oxidation of isoprene and terpenes is considered to be a very important source of formic acid^{16,17}, although some uncertainty still exists. During our experiments a clear diurnal variation of the concentration of formic acid was observed, with the highest values during the afternoon (hours

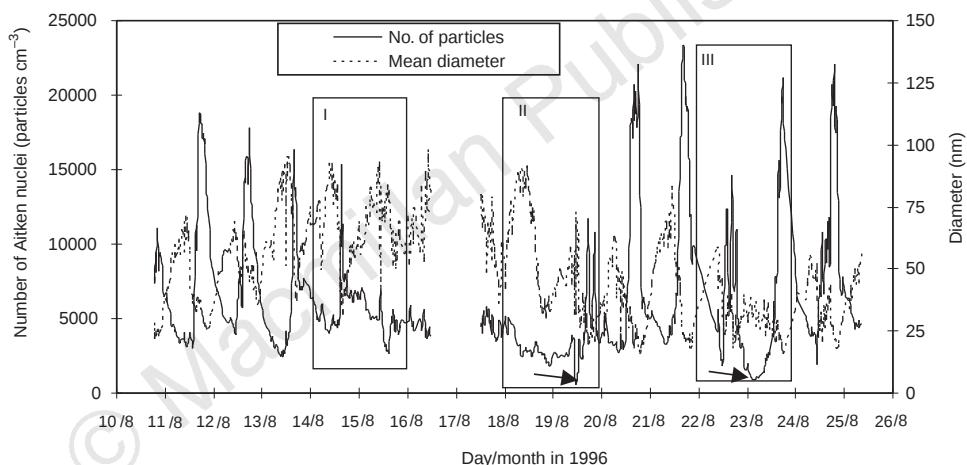


Figure 1 Variation of Aitken nuclei concentration and their mean diameter during the experiment. The three intensive measurement periods (I, II and III) are shown; arrows indicate rain events.

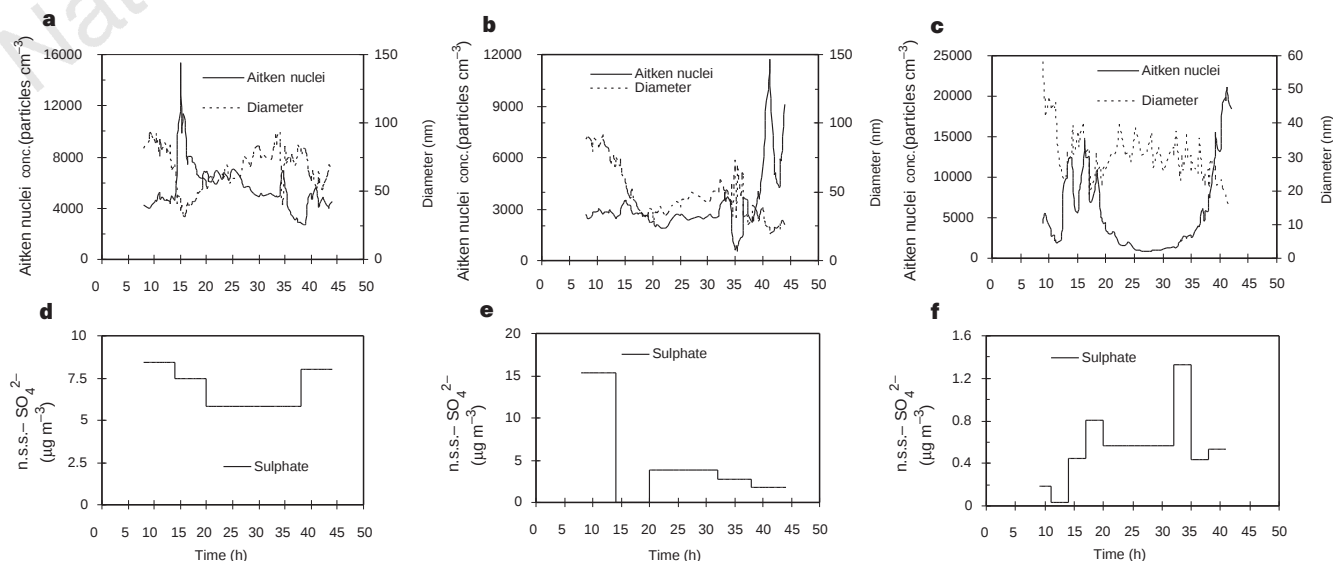


Figure 2 Data on Aitken nuclei and non-seasalt sulphate obtained during intensive measurement periods I, II and III. Shown are the variation of Aitken nuclei concentration and their mean diameter (a) (I), b (II) and c (III)), and the

variation of non-seasalt sulphate (n.s.s.-SO₄²⁻) concentration (d) (I), e (II) and f (III)). Note that the concentration scales are different for the different plots.

Table 1 Aerosol data

Time (h)	Concentrations (ng m ⁻³)						Contribution (%)		
	Pinonaldehyde	Nopinone	Wax <i>n</i> -alkanals	<i>cis</i> -Pinonic acid	<i>trans</i> -Pinonic acid	Wax <i>n</i> -alkanoic acids	<i>cis</i> -Pinonic acid	<i>trans</i> -Pinonic acid	Total pinonic acid
First intensive period (I)									
8–14	8.77	2.36	0.33	13.80	5.45	14.13	1.04	0.41	1.45
14–20	5.63	1.48	2.24	60.09	25.78	21.07	14.29	6.13	20.42
20–32	32.12	13.24	1.47	11.13	1.51	25.23	1.23	0.17	1.40
32–38	4.07	9.01	0.00	51.11	24.56	6.17	4.91	2.36	7.26
38–44	0.17	0.29	2.19	97.74	42.92	11.08	16.28	7.15	23.42
Second intensive period (II)									
8–14	0.68	0.09	0.00	11.25	4.68	1.73	1.47	0.61	2.08
14–20	1.48	0.00	0.00	16.94	10.14	0.75	11.48	6.87	18.35
20–32	6.70	0.50	0.00	7.05	2.04	12.43	8.51	2.46	10.97
32–38	1.65	0.03	0.00	12.16	5.23	16.70	9.03	3.88	12.92
38–44	0.49	0.20	0.00	20.70	8.18	10.87	29.03	11.48	40.50

Measured aerosol component concentration. Also shown are the contributions of *cis*- and *trans*-pinonic acid, and total pinonic acids, to the total mass of condensation nuclei.

14–20 and 38–44; Fig. 3a for period I and Fig. 3b for period II) and the lowest during the night (hours 20–32; Fig. 3a and b), indicating that formic acid is produced by photo-oxidation. This diurnal variation is also in agreement with the corresponding variations in the concentrations of ozone³ (60–70 parts per billion, p.p.b.v., during daytime and below 10 p.p.b.v. at night) and Aitken nuclei (Fig. 2a and b). These observations suggest a common origin for gaseous formic acid and new particle formation, that is, photo-oxidation of natural hydrocarbons (isoprene and terpenes) emitted from the forest. Isoprene concentrations measured during the same experiment showed a similar pattern, which is in agreement with those observed for Aitken nuclei and formic acid³. In addition, the ratio of the concentrations of formic to acetic acid was ~1.5, a value that is considered to be characteristic of growing vegetation¹⁷. Although it is known that formic acid and acetic acid are not able to produce new particles, as they are too volatile¹⁶, the co-variation of the concentrations of formic acid and of Aitken nuclei can be taken as strong evidence for the rapid photo-oxidation reactions involving isoprene and terpenes. The diurnal variation of particulate *cis*- and *trans*-pinonic acid concentration follows exactly the same pattern as the corresponding variation of formic acid and Aitken nuclei concentration. Considering that the higher ozone concentration³ is associated with low α -pinene and β -pinene concentrations (their concentrations range from a few parts per trillion by volume, p.p.t.v., during daytime up to 3 p.p.b.v. during night)³ we can

further conclude that these two acids (associated with particles <500 nm) are photo-oxidation products of α -pinene, leading further to new particles that are formed over the forest.

The variation of the concentrations of pinonaldehyde and nopinone presents a diurnal pattern that is inverse to those of Aitken nuclei, formic acid, and *cis*- and *trans*-pinonic acid. The low pinonaldehyde and nopinone concentrations in daytime could be explained by their photolysis¹⁸ (lifetime of 3.3 hours for pinonaldehyde) and/or reaction with OH radicals^{18,19}. On the other hand, their increase during the night could result from reactions of α - to β -pinene with NO₃ and/or O₃. The presence of these two carbonyl compounds in the particulate matter, despite their high vapour pressure (5.1 Pa at 298 K for pinonaldehyde¹⁸), can be explained by the assumption that vapour-phase products condense onto existing seed particles. In addition, once organic compounds have begun to condense and an organic layer has formed on the particles, even products whose gas-phase concentrations are below their saturation concentrations will partition a portion of their mass into this condensed organic phase²⁰. Lower temperatures that occur during the night would favour the above process²⁰.

Our results indicate that the observed formation of new particles over forests is most probably caused by interaction of organic acids, for example, pinonic acid, produced by the photo-oxidation of terpenes with other organic or inorganic species present in the atmosphere. □

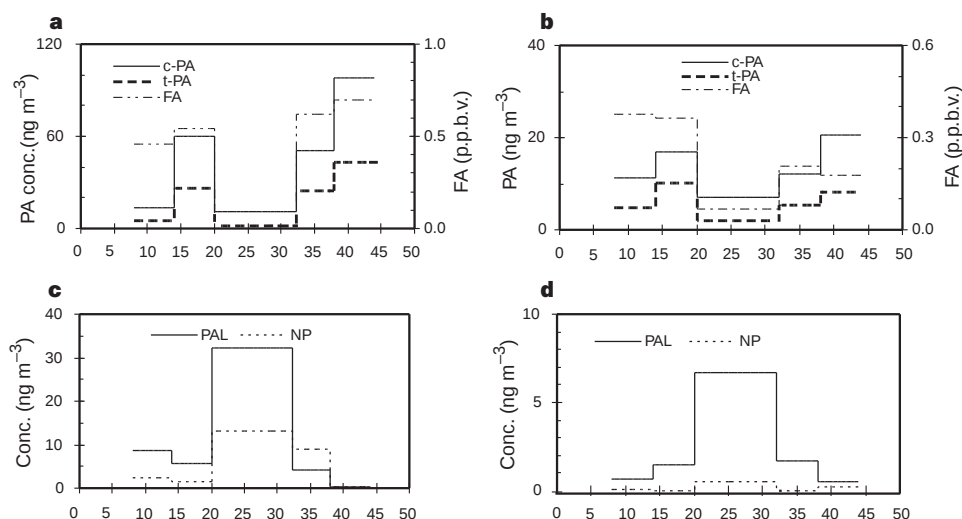


Figure 3 Data on organic compound concentrations during intensive measurement periods I and II. Shown are variation of formic acid (FA), *cis*- (c-PA) and *trans*-pinonic acid (t-PA) (**a** (I) and **b** (II)), pinonaldehyde (PAL), and nopinone (NP)

concentrations (**c** (I) and **d** (II)) (see Fig. 2a, b). Note that the concentration scales are different for the different plots.

Methods

Measurements of the concentration of Aitken nuclei and their size distribution in the range 3–200 nm were made every 15 min, by using an ultrafine Aitken nuclei counter (TSI 3025A), equipped with a diffusion battery (TSI 3040). For the identification of organic compounds, aerosols were collected on pre-cleaned (550 °C, overnight) glass-fibre filters (collection efficiency > 99%) by using a high-volume sampler (flow rate 60–75 m³ h⁻¹). Samples were collected every 6 h during sunlight and every 12 h during night-time for two sampling periods of 48 h each. Size-distributed aerosol samples were collected at three 12-h intervals (08:00–20:00, 20:00–08:00 and 08:00–20:00 local time) during the third intensive sampling period by using a five-stage (plus back-up filter) Sierra Andersen Model 230 impactor. Aerosol particles were separated into six size fractions, according to the following equivalent cut-off diameters at 50% efficiency: 7.2, 3.0, 1.5, 0.96, 0.5 µm and back-up filter < 0.5 µm. To eliminate the possibility of artefacts, tests were performed by using: (1) the sampling device alone, and (2) the same sampling device with the addition in front of it of a denuder system containing 900 tubes (each tube has a section of 4 × 4 mm and a height of 20 cm) coated with a water/glycerol KNO₃ solution in order to remove O₃ and other oxidants^{21,22}. No significant differences were observed between the two devices regarding the concentrations of pinonaldehyde (5.2–9.0%), nopinone (5.8–11.0%) and of *cis*- (0.2–1.0%) and *trans*-pinonic (1.4–2.6%) acid²². A detailed description of the analytical procedure used for extraction, separation and analysis of the main organic compounds has been published elsewhere²³. The separated organic compounds were analysed by high-resolution gas chromatography and gas chromatography-mass spectrometry in the electron impact and chemical ionization mode. Gaseous formic and acetic acids, and aerosol with a equivalent cut-off diameter < 2.0 µm, were collected by using an annual denuder equipped with a filter pack system²⁴. The analyses of the denuder and filters extracts were performed by ion chromatography⁵.

Received 29 December 1997; accepted 4 August 1998.

- Novakov, T. & Penner, J. E. Large contribution of organic aerosols to cloud-condensation nuclei concentrations. *Nature* **365**, 823–826 (1993).
- Fehsenfeld, F. *et al.* Emissions of volatile organic compounds from vegetation and the implications for atmospheric chemistry. *Glob. Biogeochem. Cycles* **6**, 389–430 (1992).
- Bonsang, B. *et al.* Aerosols Formation from Biogenic Organic Carbon (Rep. No. 1-CT PL 951352, Environment and Climate, European Commission—DG XII, Brussels, 1997).
- Mäkelä, J. M. *et al.* Observations of ultrafine aerosol particle formation and growth in boreal forest. *J. Geophys. Res.* **24**, 1219–1222 (1997).
- Mihalopoulos, N., Stephanou, E. G., Kanakidou, M., Pilatzidis, S. & Bousquet, P. Tropospheric aerosol ionic composition in the Eastern Mediterranean region. *Tellus B* **49**, 314–326 (1997).
- Hakola, H., Arey, J., Aschmann, S. M. & Atkinson, R. Product formation from the gas-phase reactions of OH radicals and O₃ with a series of monoterpenes. *J. Atmos. Chem.* **18**, 75–102 (1994).
- Wilson, Wm. E., Schwartz, W. E. & Kinzer, G. W. Haze Formation—its Nature and Origin (US Environmental Protection Agency MD-52, Research Triangle Park, NC, 1972).
- Gogou, A., Stratigakos, N., Kanakidou, M. & Stephanou, E. G. Organic aerosols in Eastern Mediterranean: components source reconciliation by using molecular markers and atmospheric back trajectories. *Org. Geochem.* **25**, 79–96 (1996).
- Grosjean, D., Williams, E. L. & Seinfeld, J. H. Atmospheric oxidation of selected terpenes and related carbonyls: gas-phase carbonyl products. *Environ. Sci. Technol.* **26**, 1526–1533 (1992).
- Yokouchi, Y. & Ambe, Y. Aerosols formed from the chemical reactions of monoterpenes and ozone. *Atmos. Environ.* **19**, 1271–1276 (1985).
- Atkinson, R. Gas-phase tropospheric chemistry of organic compounds. *J. Phys. Chem. Ref. Data. Monogr.* **2**, 1–216 (1994).
- Hatakeyama, S., Izumi, K., Fukuyama, T. & Akimoto, H. Reactions of ozone with α -pinene and β -pinene in air: yields of gaseous and particulate products. *J. Geophys. Res.* **94**, 13013–13024 (1989).
- Hoffmann, T. *et al.* Formation of organic aerosols from the oxidation of biogenic hydrocarbons. *J. Atmos. Chem.* **26**, 189–222 (1997).
- Waengberg, I., Barnes, I. & Becker, K. H. Product and mechanistic study of the reaction of NO₃ radical with α -pinene. *Environ. Sci. Technol.* **31**, 2130–2135 (1997).
- Hatakeyama, S., Izumi, K., Fukuyama, T., Akimoto, H. & Washida, N. Reactions of OH with α -pinene and β -pinene in air: estimate of global CO production from the atmospheric oxidation of terpenes. *J. Geophys. Res.* **96**, 947–958 (1991).
- Keene, W. C. & Galloway, J. N. Considerations regarding sources for formic and acetic acids in the troposphere. *J. Geophys. Res.* **91**, 14466–14474 (1986).
- Jacob, P. J. & Wofsy, S. J. Photochemistry of biogenic emissions over the Amazon forest. *J. Geophys. Res.* **93**, 1477–1486 (1988).
- Hallquist, M., Wangberg, I. & Ljungstrom, E. Atmospheric fate of carbonyl oxidation products originating from α -pinene and Δ^3 -carene: Determination of rate of reaction with OH and NO₃ radicals, UV absorption cross sections, and vapor pressures. *Environ. Sci. Technol.* **31**, 3166–3172 (1997).
- Atkinson, R. & Aschmann, S. M. Atmospheric chemistry of the monoterpene reaction products nopinone, camphenilone and 4-acetyl-1-methylcyclohexene. *J. Atmos. Chem.* **16**, 337–348 (1993).
- Odum, J. *et al.* Gas/particle partitioning and secondary organic aerosol yields. *Environ. Sci. Technol.* **30**, 2580–2585 (1996).
- Geyh, A., Wolfson, J., Koutrakis, P. & Mulic, J. Development of an Active Personal O₃ Sampler Using a Hollow Tube Diffusion Denuder (US Environmental Protection Agency/600A-94/154, Research Triangle Park, NC, 1994).
- Kavouras, I. *In-situ* Study of the Mechanism of Formation of Secondary Organic Aerosols through the Photooxidation of Monoterpenes. Thesis, Univ. Crete (1998).

- Gogou, A. I., Apostolaki, M. & Stephanou, E. G. Determination of organic molecular markers in marine aerosols and sediments: one-step flash chromatography compound class fractionation and capillary gas chromatographic analysis. *J. Chromatogr. A* **799**, 215–231 (1998).
- Lawrence, J. & Koutrakis, P. Measurement and speciation of gas and particulate phase organic acidity in an urban environment. I. Analytical. *J. Geophys. Res.* **101**, 9159–9163 (1996).

Acknowledgements. We thank M. Apostolaki and A. Lemonakis for technical assistance. This work was supported by the European Commission DGXII.

Correspondence and requests for materials should be addressed to E.G.S. (e-mail: stephanou@chemistry.uch.gr).

Mantle convection simulations with rheologies that generate plate-like behaviour

Ron Trompert* & Ulrich Hansen†

* Faculty of Earth Sciences, Utrecht University, PO Box 80021, NL-3508 TA Utrecht, The Netherlands

† Institut für Geophysik, Westfälische Wilhelms Universität Münster, Corrensstrasse 24, D-48149 Münster, Germany

A long-standing problem in geodynamics is how to incorporate surface plates in numerical models of mantle convection. Plates have usually been inserted explicitly in convection models as rigid rafts^{1–3}, as a separate rheological layer^{4,5} or as a high-viscosity region within weak zones^{6–11}. Plates have also been generated intrinsically through the use of a more complex (non-newtonian) rheology for the entire model^{12,13} but with a prescribed mantle flow. However, previous attempts to generate plates intrinsically and in a self-consistent manner (without prescribed flow) have not produced surface motions that appear plate-like^{14–16}. Here we present a three-dimensional convection model that generates plates in a self-consistent manner through the use of a rheology that is temperature and strain-rate dependent, and which incorporates the concept of a yield stress. This rheology induces a stiff layer on top of a convecting fluid, and we find that this layer breaks at sufficiently high stresses. The model produces a style of convection that contains some of the important features of plate tectonics, such as the subduction of the stiff layer and plate-like motion on the surface of the fluid mantle. However, the model also produces some non-Earth-like features, such as episodic subduction followed by the slow growth of a new stiff layer, which may be more consistent with the style of convection found on Venus.

It is known that for temperature-dependent viscosity convection with a viscosity contrast larger than 10⁴–10⁵, a cold stagnant lid develops at the top of the fluid layer with a hot convecting layer underneath^{17–19}. In this stagnant lid, heat transport is strongly dominated by conduction, and the largest viscosity variations occur in the cold boundary layer. This stagnant lid can be mobilized by weak zones owing to failure of the lid at sufficiently high stresses²⁰. We have implemented a rheology in a three-dimensional mantle convection model that is temperature-dependent, to produce a stagnant lid, and that facilitates ductile failure through a strain-rate-dependent part. The model we have used is fully self-consistent; this means that the plate-like behaviour of the fluid at the top—including the development of plate-like regions—comes from the combination of the chosen rheology and the convective processes taking place. No special boundary conditions are imposed, nor have we specified different rheologies for different horizontal fluid layers. No assumptions were made with respect to the pre-existence of plates and plate geometry.

We have considered thermally driven convection in an extremely viscous Boussinesq fluid. Flow of this type is described by the Stokes equations. (For a detailed description regarding the numerical method we have used to solve these equations, see refs 21, 22.)

The rheology that we have adopted is based on the one used in ref. 23, and is related to the Bingham model which states that a material is rigid until a yield stress is reached. For stresses larger than this yield stress, failure occurs and the material behaves like a newtonian fluid.

The non-dimensional viscosity is defined by the following law:

$$\eta = \frac{2}{\frac{1}{\eta_T} + \frac{1}{\eta_e}}, \eta_T = \exp(-RT), \eta_e = \eta^* + \frac{\sigma}{(\bar{\epsilon} : \bar{\epsilon})^{1/2}}$$

where R is a constant governing the viscosity contrast based on temperature T , and $\bar{\epsilon}$ is the strain-rate tensor. The parameters η^* and σ^* can be chosen freely. However, they need to be chosen carefully as σ^* controls the stress level at which failure occurs, and η^* controls the behaviour of the fluid for higher stresses. If the yield stress is very low, the stagnant lid yields almost everywhere and it would not consist of large rigid regions surrounded by localized regions of high

deformation as on Earth. On the other hand, when the yield stress is very high, the lid stays intact, which is not very Earth-like either. This means that in order to approach the situation on Earth as closely as possible, these parameters should be chosen so that the yield stress is low enough for yielding to occur and the regions where yielding takes place are very localized.

We have done numerical convection simulations in a $4 \times 4 \times 1$ bottom-heated cartesian box with stress-free impermeable boundaries. This translates into a domain (in km) of $11,600 \times 11,600 \times 2,900$ when we take the height of the box equal to the depth of the mantle. The Rayleigh number is a measure of the ratio of buoyancy forces over viscous forces, and its value based on the temperature-dependent viscosity at the top is 100. A $65 \times 64 \times 32$ grid was used; the horizontal grid is uniform and the grid in the vertical is non-uniform with refinement in the thermal boundary layers. The simulation was performed over a period of 0.23 non-dimensional time units or 60 Gyr, assuming a length scale equal to the height of the mantle and a heat diffusion coefficient of $10^{-6} \text{ m}^2 \text{ s}^{-1}$. The temperature-based viscosity contrast over the box is 10^5 ($R = 5 \log(10)$). The computation was started from a fully developed stagnant-lid convection. We have chosen the values of σ^* and η^* by taking into account the considerations above. These values were determined from preliminary runs, and were chosen to be 2.0 and 0.00001 for σ^* and η^* , respectively.

The style of convection has an episodic character. It is characterized by cycles consisting of failure in regions of the stagnant lid and subduction of parts of the lid on a short timescale, followed by a relatively long period of growth of a new stagnant lid and subduction of parts of the lid on a short timescale, followed by a relatively long period of growth of a new stagnant lid where the old one has subducted. Figure 1 shows such a cycle by means of 'snapshots' of the temperature at six different times. Figure 2 shows the temporal evolution of the Nusselt number and surface toroidal-poloidal kinetic energy ratio during this cycle. The Nusselt number is the non-dimensional vertical heat flux averaged over the volume, and measures the heat transport efficiency of the flow. A toroidal velocity field possesses only a vertical vorticity component, whereas a poloidal field has only horizontal vorticity components.

On failure, a ridge-like structure develops with an active upwelling flow underneath and at the same time downward flow is initiated (Fig. 1a, b). The part of the cold lid between the ridge and the subduction zone accelerates very rapidly towards the downwelling region. This process is accompanied by a strong increase in Nusselt number and in surface toroidal-poloidal kinetic energy ratio (Fig. 2). For this particular cycle, the toroidal-poloidal kinetic energy ratio has a maximal value of ~ 0.19 . This ratio is significant, but is small compared with the value of approximately unity that is observed on Earth. During this phase, the part of the lid between the ridge and the subduction zone moves as a plate, that is, it shows an almost constant velocity over the moving region and simultaneously sharp velocity changes at its boundaries. A close-up view of such a moving part of the lid during another (but similar) cycle is shown in Fig. 3, and brings out the features of a sharply varying velocity near the ridge, a somewhat smoother velocity change in regions of transform motion and a virtually constant velocity within the region. Translated to the Earth, we obtain a plate speed of 4 cm yr^{-1} , which is a realistic order of magnitude. The cold lid descends into the interior as a sheet-like downwelling (Fig. 4) which resembles the geometry of observed subduction zones. Eventually the lid collides with, and spreads out over, the bottom boundary, after which a trench migration phenomenon occurs (Fig. 1c–e). This trench migration appeared several times during the whole simulation and appears to be typical. For Earth-like circumstances, the velocity associated with the trench migration is of the order of 7 cm yr^{-1} , which is again close to the observed values. The subducted lid remains at the bottom of the box. A new stagnant lid grows where the old one has subducted the subducted lid slowly

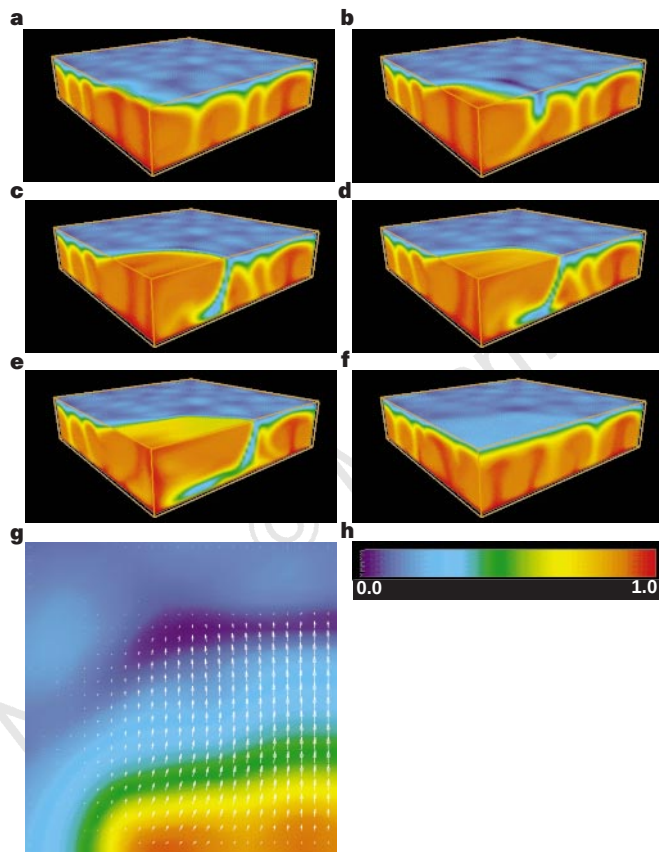


Figure 1 A typical convection cycle. Here the temperature is shown; the colour coding is indicated in **h**, where red is hot and blue is cold. The top surface of the box has been removed. The time points associated with the panels **a–f** are indicated in Fig. 2. After failure, a ridge-like structure develops with an upwelling flow underneath. This is shown in **a**, where the yellowish colour in the corner closest to the observer is the ridge. A part of the blue stagnant lid starts to move in the direction of the corner which is at the right side of the panel. Subduction of this moving part is initiated in the region where the flow goes downwards and only the part of the lid between the ridge and the down flow is in motion. It moves as a plate with little internal strain (**b**); the velocity field of this moving part of **b** clearly demonstrates this (**g**). The maximum velocity is 3.2 cm yr^{-1} ; **c, d**, show that the subducted part of the lid collides with the bottom boundary and spreads out. At this time a trench migration occurs. The flow initiated by the subduction is now pushing back the trench where subduction took place (**c–e**). After this, the slab shown in the **c–e** will now disconnect from the stagnant lid on top and will stay at the bottom of the box. This slab will heat up and at the same time the subducted part of the lid will be replaced by a new lid through cooling from the top (**f**).

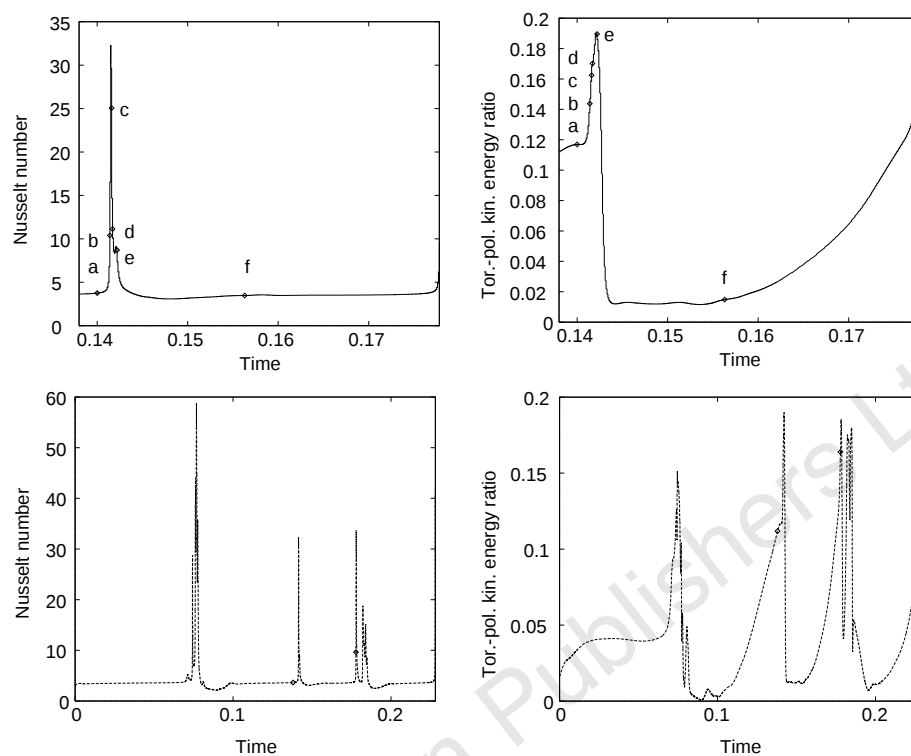


Figure 2 Time history plot of Nusselt number and surface toroidal-poloidal energy ratio. The Nusselt number as a function of time (left) and the surface toroidal-poloidal kinetic energy ratio (right) for the cycle shown in Fig. 1 (upper) and for the

whole simulation (lower). Times indicated by a-f in the upper panels correspond to the panels shown in Fig. 1. The time span of the upper two panels is indicated by the dots in the lower panels.

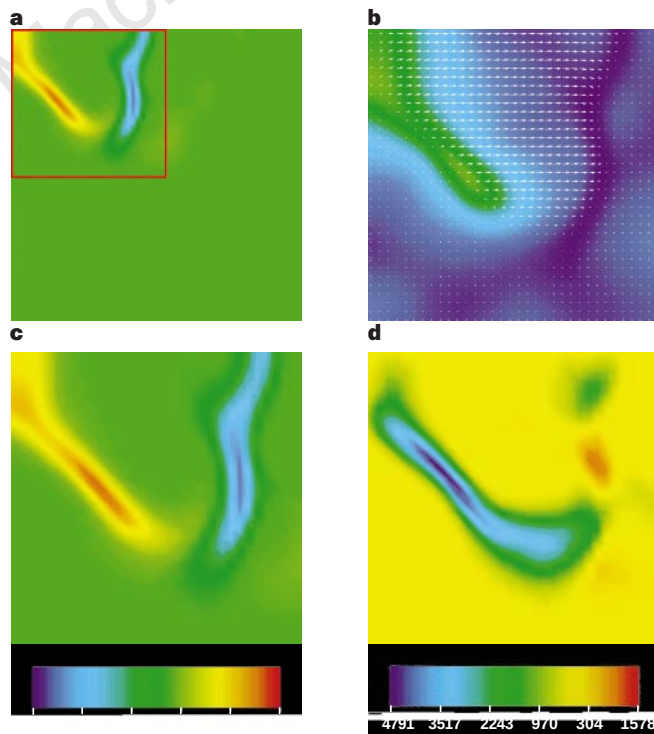


Figure 3 An example of plate-like motion. This figure is taken from a different cycle from that in Fig. 1 and, unlike Fig. 1, features a divergent margin with a skewed part that does not coincide with the boundary. In **a**, the top surface of the whole box is shown to indicate the plate-like region. The colours indicate the horizontal divergence. The other panels show a close-up view of the region indicated by the red square. **b**, Velocity vectors. The colours indicate the temperatures slightly below the upper surface of the box where blue is cold

and green is hot. The velocity vectors show no internal strain. The largest velocity vector corresponds to a velocity of 1.1 cm yr^{-1} . The horizontal divergence and vertical vorticity are displayed in **c** and **d**, respectively. The colour bars indicate the non-dimensional values. The convergent and divergent margins are identified by negative and positive values of the horizontal divergence, respectively (**c**). **d**, A very localized amount of high vertical vorticity coincides with the skewed part of the divergent margin.

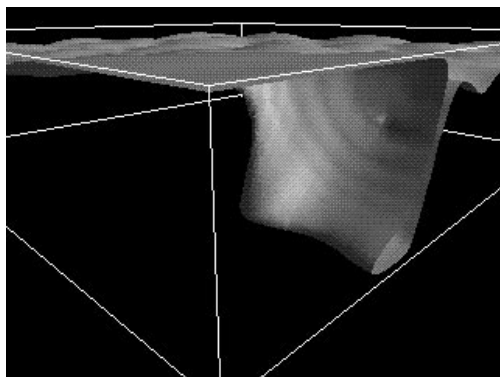


Figure 4 An example of sheet-like downwelling flow, showing a temperature isosurface.

heats up, until the stagnant-lid style of convection is reached again (Fig. 1e–f).

This cycle repeated itself several times; this is shown in Fig. 2 where the temporal evolution of the Nusselt number and of the surface toroidal-poloidal kinetic energy ratio is displayed for the whole simulation. The narrow peaks correspond to a failure of the stagnant lid. The subduction process occurs on a timescale that is shorter than the time needed for a new lid to grow to its full thickness, which is about an order of magnitude larger for this simulation. This is indicated in Fig. 2, where the subduction process stops at point e and the subducted lid is replaced at point f. The formation of ridges did not only occur near the boundaries of the box, as one might expect looking at Fig. 1, but also in the middle of the box. During the simulation, the lid subducted usually by parts, but we have also seen one event of an almost complete resurfacing.

Our experiment shows several features related to plate tectonics as currently observed. The rheology produces plate-like regions with very little internal strain, fairly sharp changes in velocity at plate boundaries, downwelling flow that is sheet-like, and regularly occurring trench migration. The surface toroidal-poloidal kinetic energy ratio is significant, but it is low compared with that of the Earth. Including a mechanism of self-lubrication into the rheology^{12,13} may increase this ratio. Further, the Earth does not currently exhibit tectonic activity with a strong episodic character. The process of subduction followed by a long period of growth of a new stagnant lid is more similar to the process that is active on Venus²⁴. When the lid fails, in the current model, it is removed much faster by subduction than it can be replaced by new material that is cooled from the top. This episodic behaviour would not occur when the rate of disappearance and the rate of replacement of the lid are the same. Then one would obtain a continuous process consisting of failure, subduction and replacement taking place simultaneously. Such a process is believed to be active on Earth.

We note that our model is still simple, and that many features—such as phase boundaries, internal heating and the presence of chemically buoyant continents—are not included. Our model needs to be modified by including ‘memory’ into the rheology as plate boundaries may continue to exist even when they are not active. Including these effects could bring the behaviour of our model closer to plate tectonics as it is observed on Earth. Our calculations, however, do show that an appropriate rheology translates the internal dynamics into plate-like surface motion. In our view, this is an important step towards a unifying model of mantle convection and plate tectonics. □

Received 8 September 1997; accepted 14 July 1998.

1. Gable, C. W., O’Connell, R. J. & Travis, B. J. Convection in three dimensions with surface plates: Generation of toroidal flow. *J. Geophys. Res.* **96**, 8391–8405 (1991).
2. Hager, B. H. & O’Connell, R. J. A simple global model of plate dynamics and mantle convection. *J. Geophys. Res.* **86**, 4843–4867 (1981).
3. Olson, P. & Corcos, G. M. A boundary layer model for mantle convection with surface plates. *Geophys. J. R. Astron. Soc.* **62**, 195–219 (1980).

4. Weinstein, S. A. & Olson, P. Thermal convection with non-Newtonian plates. *Geophys. J. Int.* **111**, 515–530 (1992).
5. Ribe, N. M. The dynamics of thin shells with variable viscosity and the origin of toroidal flow in the mantle. *Geophys. J. Int.* **110**, 537–552 (1992).
6. Davies, G. F. Mantle convection model with a dynamic plate: Topography, heat flow and gravity anomalies. *Geophys. J. Int.* **98**, 461–464 (1989).
7. Gurnis, M. Large-scale mantle convection and the aggregation and dispersal of supercontinents. *Nature* **332**, 695–699 (1988).
8. King, S. D. & Hager, B. H. The relationship between the plate velocity and trench viscosity in Newtonian and power law subduction of calculations. *Geophys. Res. Lett.* **17**, 2409–2412 (1990).
9. Schmeling, H. & Jacoby, W. R. On modelling the lithosphere in mantle convection with non-linear rheology. *J. Geophys.* **50**, 89–100 (1981).
10. Zhong, S. & Gurnis, M. Towards a realistic simulation of plate margins in mantle convection. *Geophys. Res. Lett.* **22**, 981–984 (1995).
11. Zhong, S. & Gurnis, M. Interaction of weak faults and non-Newtonian rheology produces plate tectonics in a 3D model of mantle flow. *Nature* **383**, 245–247 (1996).
12. Bercovici, D. A simple model of plate generation from mantle flow. *Geophys. J. Int.* **114**, 635–650 (1993).
13. Bercovici, D. Plate generation in a simple model of lithosphere-mantle flow with dynamic self-lubrication. *Earth. Planet. Sci. Lett.* **144**, 41–51 (1996).
14. Christensen, U. Convection in a variable viscosity fluid: Newtonian versus power law. *Earth. Planet. Sci. Lett.* **64**, 153–162 (1983).
15. Christensen, U. & Harder, H. 3D convection with variable viscosity. *Geophys. J. Int.* **104**, 213–226 (1991).
16. Čadež, O., Richard, Y., Martinec, Z. & Matyska, C. Comparison between Newtonian and non-Newtonian flow driven by internal loads. *Geophys. J. Int.* **112**, 103–114 (1993).
17. Nataf, H. C. & Richter, F. M. Convection experiments in fluids with highly temperature dependent viscosity and the thermal evolution of the Earth. *Phys. Earth Planet. Inter.* **29**, 320–329 (1982).
18. Solomatov, V. S. Scaling of temperature-dependent and stress-dependent viscosity convection. *Phys. Fluids* **7**, 266–274 (1995).
19. Stengel, K. C., Oliver, D. C. & Booker, J. R. Onset of convection in variable-viscosity fluid. *J. Fluid Mech.* **120**, 411–431 (1982).
20. Fowler, A. C. Boundary layer theory and subduction. *J. Geophys. Res.* **98**, 21997–22005 (1993).
21. Trompert, R. A. & Hansen, U. The application of a finite volume multigrid method to 3D flow problems in a highly viscous fluid with a variable viscosity. *Geophys. Astrophys. Fluid Dyn.* **83**, 261–291 (1996).
22. Trompert, R. A. & Hansen, U. On the Rayleigh number dependence of convection with a strongly temperature-dependent viscosity. *Phys. Fluids* **10**, 351–360 (1998).
23. Burgess, S. L. & Wilson, S. D. R. Spin-coating of a viscoplastic material. *Phys. Fluids* **8**, 2291–2297 (1996).
24. Head, J. W. Venus after the flood. *Nature* **372**, 729–730 (1994).

Acknowledgements. We thank D. Bercovici and C. W. Gable for comments and suggestions, and the Dutch national computing facilities foundation (NCF) for computer time on a Cray C916. This work was supported by the Dutch National Science Foundation NWO.

Correspondence and requests for materials should be addressed to U.H. (e-mail: hansen@earth.uni-muenster.de).

Two-dimensional matches from one-dimensional stimulus components in human stereopsis

Bart Farell

Institute for Sensory Research, Syracuse University, Syracuse, New York 13244-5290, USA

Center for Vision Research, SUNY Health Science Center, Syracuse, New York 13210-2375, USA

Three-dimensional visual scenes project onto the retina of the eye as two-dimensional images. The third dimension, depth, is projected as subtle differences between left and right retinal images. As early as the 1830s, stereoscopic depth perception was shown to depend on horizontal disparities between these images¹. To detect disparity, the visual system must match corresponding parts of the two retinal images. To identify the stimulus elements used in stereo matching, I applied a disparity-adaptation technique to visual patterns whose one-dimensional components and two-dimensional features have very different disparities. Surprisingly, the adaptors that are effective in altering depth perception appear widely separated in depth from the patterns they adapt. I conclude that stereo matching occurs in all directions of two-dimensional space and that one-dimensional components are the stimulus primitives, the fundamental elements of stereo matching. This is a reversal of the classical view of stereo correspondence as a one-

dimensional (horizontal) matching of monocular two-dimensional features²⁻⁴.

Figure 1a shows a stereogram containing a horizontally offset line and a zero-disparity occluder. When the half-images are fused (by crossing one's eyes), the line is revealed to be behind the plane of fixation, just as if no occluder were present. However, the superimposed half-images (Fig. 1b) show a variety of local disparity directions between the segments of the occluder, including horizontal, vertical and oblique directions. These 'aperture disparities' have a well-known analogue in motion perception⁵, but are absent from the non-overlapping stimuli typically used to study stereopsis. Thus, in naturalistic, 'layered' scenes of occluding and transparent surfaces, stereo correspondence becomes a two-dimensional (2D) matching problem in which horizontal disparity is not a reliable cue to depth. This dissociation of retinal disparity and perceived depth can be exploited to identify the stimulus primitives of human stereopsis.

Here I use 'stereo plaids', the simplest frequency-domain 2D stereo stimuli, which are made by superimposing two sinusoidal gratings (Fig. 1c). Unlike broad-band, sharp-edged patterns (Fig. 1a), sinusoidal stereo plaids do not segregate in depth, even if their components have very different disparities. Instead, they cohere in depth, forming a 2D pattern within a single depth plane^{6,7}. The disparities of stereo plaids are, like aperture disparities, generally non-horizontal. Depth coherence has a familiar one-dimensional (1D) analogue in the perception of a squarewave grating (alternating light and dark stripes) or similar broadband pattern. One sees a unified grating at the depth corresponding to the disparity of the fundamental, despite the differing disparities of the higher harmonics. The simplest explanation is that the visual system matches edges of the squarewave—and 2D features of plaids—not the harmonics.

To identify the matching primitives, it is necessary first to determine how the perceived depth of a plaid is related to the disparities of its components. I presented superimposed gratings

orientated at 15° and 45° to observers, who classified the resulting plaids as 'near' or 'far' with respect to the plane of fixation. The disparity of one grating was zero and that of the other, expressed as a phase angle, ranged over $\pm\pi/4$. Stimuli appeared briefly, to preclude eye movements that could alter the disparities. When the 15° grating had zero disparity, the plaid was seen as 'near' if the 45° grating had negative disparity and as 'far' if it had positive disparity. However, when the 45° grating had zero disparity, the plaid was seen with reversed depth: as 'far' if the disparity of the 15° grating was negative and as 'near' if it was positive (Fig. 2). It is clear that relative depth is not predictable from component disparities alone (and that depth coherence is not depth capture⁸ or disparity averaging⁹).

What does predict the depth judgements of Fig. 2 is the horizontal component of the disparity of the plaids' 2D features (for example, intersections or 'blobs'). This is given by $\rho \cos(\phi)$ for direction ϕ and magnitude ρ of the 2D feature disparity, where

$$\phi = \theta_1 - \arctan \left\{ \frac{\Delta_1 \cos(\theta_2 - \theta_1) - \Delta_2}{\Delta_1 \sin(\theta_2 - \theta_1)} \right\} \quad (1)$$

$$\rho = \left| \frac{\Delta_1}{\cos(\phi - \theta_1)} \right| \quad (2)$$

θ_1 and θ_2 are the gratings' disparity directions measured normal to their orientations, and Δ_1 (non-zero) and Δ_2 are the respective disparity amplitudes. So, depending on grating orientation, the disparities of the 2D features can be similar to, or very different from, the disparities of the component gratings. This is why adding a zero-disparity grating to a 'far' grating can yield a depth-reversed 'near' plaid.

There would seem to be little risk in concluding that the matching primitives are some set of 2D stimulus features. This hypothesis is supported by the association between the polarity (near compared with far) of perceived depth and the direction (negative compared with positive) of the horizontal component of the 2D feature

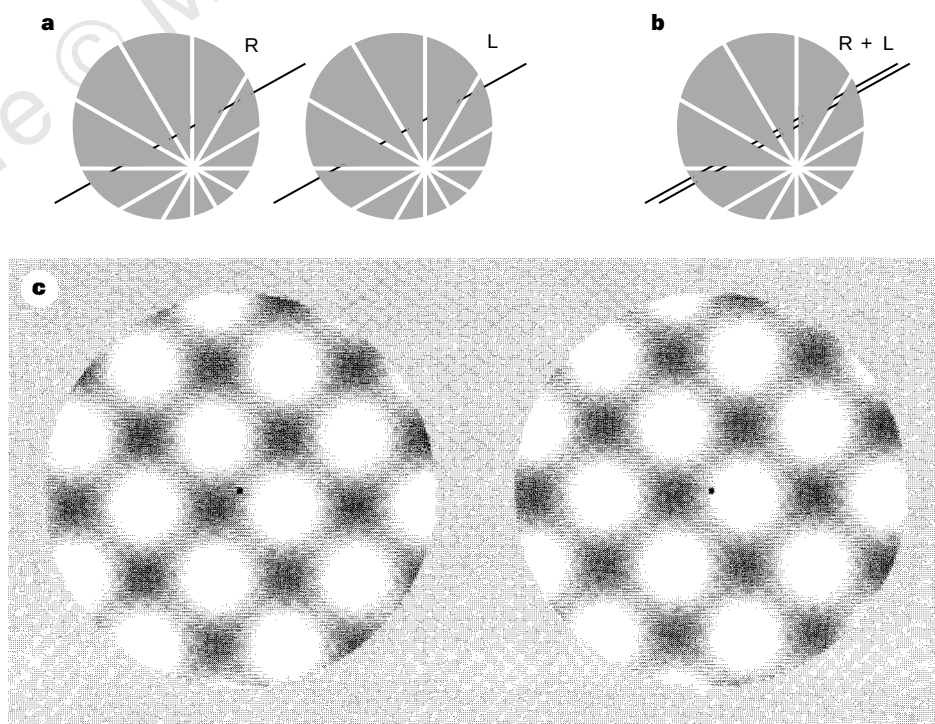


Figure 1 Stereograms. **a**, When cross-fused, the line, with positive horizontal disparity, appears behind the zero-disparity occluder. **b**, Superimposed half-images show that the line's local disparities are not confined to the horizontal meridian, but run parallel to the apertures between segments of the occluder. **c**, Stereo plaid composed of gratings with positive and zero disparity. The plaid's

2D feature disparities parallel the zero-disparity grating (135°), just as the local disparities in **a** parallel the apertures of the zero-disparity occluder. The two gratings appear at the same depth, forming a coherent plaid; observers cannot determine which grating has zero disparity.

disparity, and the lack of such an association involving the 1D component disparity (Fig. 2). In fact, however, 1D primitives cannot be excluded at this point. Next I examine and test these alternatives, taking advantage of the large disparity difference between 2D features and 1D components in depth-reversed stereo plaids.

In traditional theories, features of the 2D image are matched horizontally. However, developments such as aperture disparities, the perceived depth data of Fig. 2, and results of other studies^{10,11} show that stereo matching is not confined to the horizontal meridian or epipolar lines. A model for extracting the disparities of these features without incurring the full cost of an unconstrained 2D search uses arbitrarily oriented receptive fields (Fig. 3a). These sample left and right retinal images with a disparity direction that is normal to the orientation of the receptive field. Data must be pooled across orientations to overcome the stereo aperture problem, that is, a 1D detector cannot reliably code a 2D disparity. A pair of symmetrically oriented receptive fields yields an efficient pooling algorithm (Fig. 3a).

Under an alternative hypothesis, 1D components are the matching primitives (Fig. 3b). Physiologically realistic receptive fields such as those of Fig. 3a can be used to register matches between local 1D Fourier components of the left and right retinal images in the first of two processing stages (Fig. 3b). First-stage neural responses are combined at a second stage to compute the disparities of 2D features by implementing equations (1) and (2). The model unifies the stereo processing of 1D and 2D stimuli, matching 1D components in order to compute the 2D disparity field that the one-stage model (Fig. 3a) extracts directly. The model's second stage can also facilitate the difficult problem of perceiving transparency by selectively combining components within, but not between, broadband patterns (for example, squarewave gratings in Fig. 2). A related two-stage model of motion perception has been productive and controversial for the past 15 years¹².

These models code disparity in a 2D correspondence space and use the full range of receptive-field orientations, not just the vertical¹³ or the unorientated^{4,27}. In this respect, the receptive-field organization of both models differs significantly from models that have been proposed previously for coding the disparity of 2D stimuli. Each receptive field has its disparity direction normal to its preferred orientation and thus performs an efficient 1D match.

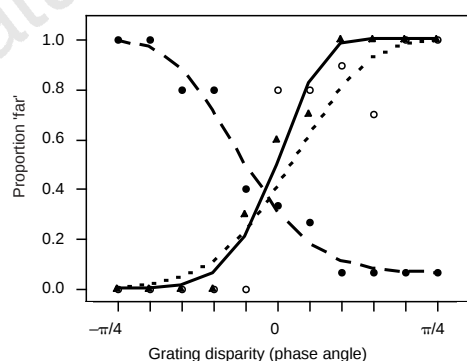


Figure 2 Perceived depth polarity of stereo plaids as a function of disparity of one component grating; the other component had zero disparity. Grating orientations were 15° and 45°. Data are for one observer; the other two observers gave similar results. Triangles, solid line: 45° grating had disparity given on abscissa. Gratings were sinusoidal and cohered in depth. Filled circles, dashed line: 15° grating had disparity given on abscissa. Gratings were sinusoidal and cohered, but now the psychometric function is reversed. Open circles, dotted line: same as filled circles, except that the gratings were squarewaves. These gratings segregate in depth; the data describe depth of the 15° grating. Depth reversal thus depends on the coherence of the gratings, and therefore on 2D stimulus structure. Lines are based on Weibull fits.

Yet disparity coding across receptive fields is fully 2D in that the correspondence search spans all directions and can establish a match in any direction. The models differ in their stereo primitives and also in their relation to feature detection. In the one-stage model, the detection of 2D monocular features occurs before stereo matching; in the two-stage model, stereo matching precedes feature detection. However, in the limiting case of a single point-like stimulus as input, the two models are indistinguishable.

Thus, we can specify the disparities of stereo plaids, and those of other 2D patterns, either as 2D feature disparities or as 1D component disparities. But which does the visual system actually use? To find an answer, I used an adaptation method. After prolonged viewing of an adapting stimulus, a spatially similar test stimulus with a neighbouring disparity appears displaced in depth away from the adaptor¹⁴. I used this displacement after-effect to enhance the discriminability of test stimuli whose disparities flank the disparity of the adaptor. Observers selected the farther of two depth-reversed plaids that differed slightly in depth (Fig. 4). If the visual system extracts the disparity of 2D features, then depth discrimination should be facilitated by adapting to disparity d_2 (adaptor A in Fig. 4), midway between the 2D feature disparities of the two plaids. If the visual system extracts the disparity of 1D components, then facilitation should result from adapting to disparity d_1 (adaptor B), midway between the 1D component disparities.

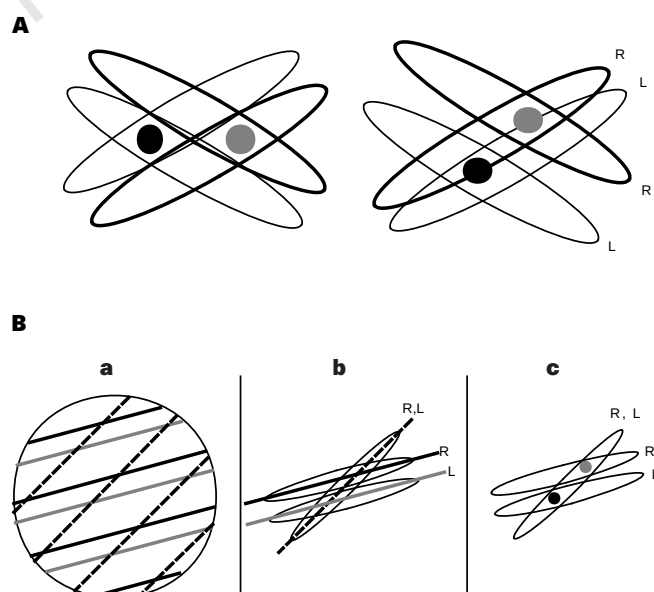


Figure 3 One- and two-stage models of stereo matching. **A**, One-stage model for coding 2D feature disparities. Receptive fields are tuned to disparities that are perpendicular to their orientations. Two such receptive fields (here showing similar spatial frequency preferences) with arbitrary orientations θ and $-\theta$ form the basic computational unit. The stimuli (disks) have horizontal (left) and oblique (right) disparities. Horizontal and vertical disparities are recoverable from the sum and difference of phase shifts, respectively: $D_H = (\phi_1 + \phi_2)(\lambda/4\pi)/\sin\theta$; $D_V = (\phi_1 - \phi_2)(\lambda/4\pi)/\cos\theta$, where ϕ_1 and ϕ_2 are the separate phase shifts and λ is the receptive-field period. **B**, Two-stage model for deriving 2D feature disparities from 1D components. **a**, A reversed-depth stereo plaid containing a 45° zero-disparity grating. **b**, First-stage coding of disparities of 1D components by neurons with elongated receptive fields. Neurons are activated to the extent that their binocular receptive fields match Fourier components of the stimulus; activation is highest for the receptive field profiles shown here. **c**, Second-stage combination of active units, implementing equations (1) and (2) to compute intersections of the receptive fields. The output is a computed disparity of 2D features (disks), the same disparity that the one-stage model codes directly (**A**, right). Note that the receptive-field disparity and the 2D feature disparity have horizontal components of opposite signs. L and R denote left and right eyes.

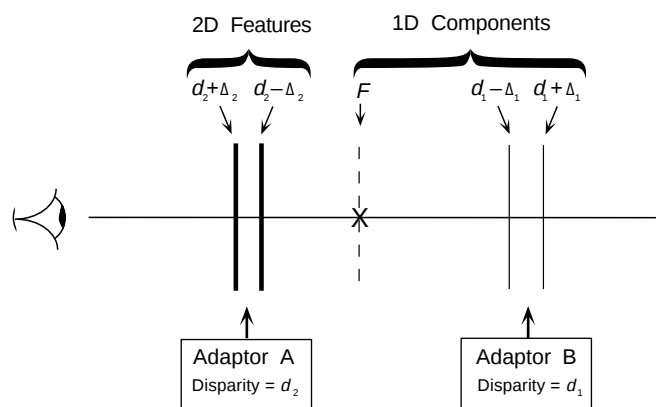


Figure 4 Adaptation experiment. Two reversed-depth stereo plaids appear in separate intervals on each trial. Observers select the interval containing the more distant of the plaids, that is, the one with disparity $d_2 - \Delta_2$. Task performance was measured before and after observers adapted to a grating or plaid whose disparity lay either between the disparities of the 2D features appearing in the two intervals (adaptor A), or between the non-zero disparities of the component gratings (adaptor B). Observers see the test stimuli as plaids in front of the fixation plane.

Figure 5 compares pre- and post-adaptation performance. Neither grating adaptors nor plaid adaptors significantly improve discrimination ($P > 0.05$) when the adapting disparity is centred between the disparities of the 2D features of the test plaids (A adaptors). However, both grating adaptors and plaid adaptors significantly improve discrimination when the adapting disparity is centred between the disparities of the 1D components (B adaptors), provided that the 1D components of adapting and test patterns are parallel. The probability of correct discrimination increased to 80% from the pre-adaptation level of 69%. Adaptors that were oriented orthogonally had no significant effect on performance. Additional conditions in which adaptors had disparities midway between disparities of the test plaids' second-order gratings¹⁵ also showed no effect. Thus, adaptation is effective at the disparity of the test plaids' 1D components, in the direction normal to these components. This occurs even though the test stimuli are not seen at the 'far' depth of the effective adaptors, but rather at the 'near' depth corresponding to the disparity of the 2D features. The results, which are predicted by the two-stage model and run counter to other models of stereo matching, implicate 1D components as the primitives of a 2D stereo correspondence process.

The visual system's use of vertical as well as horizontal disparities has been hinted at ever since von Helmholtz's observations¹⁶. A few past studies have shown that, in some conditions, naturalistic vertical disparities can influence perceived depth^{17,18}. The present results take this further, indicating that stereo matching is intrinsically a 2D correspondence process. First, disparities can have any direction during aperture viewing. Second, the two-stage model codes the stereo correspondence of 2D stimuli in two dimensions, regardless of the direction of the stimulus disparity. For example, a point stimulus whose retinal disparity is strictly horizontal would match binocular receptive fields having every correspondence direction; the correspondence would have non-zero disparity in every direction except the strictly vertical. This results from matches between elongated receptive fields and the point's isotropically distributed 1D components. It also helps to explain a long-standing physiological puzzle. Most visual cortical neurons are binocularly sensitive^{19–22}, but they do not respond selectively to horizontal disparities only. Instead, they respond to any of a variety of disparity directions. The distribution is isotropic^{23–25}, or very nearly so²⁶. This

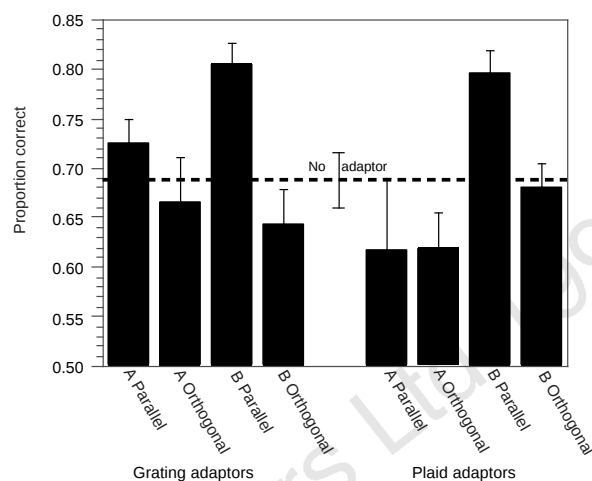


Figure 5 Depth discrimination performance. Data for one observer; the data for two other observers were similar. Chance performance is 50%. 'A' adaptors have disparities centred between those of the test patterns' 2D features (Fig. 4). 'B' adaptors have disparities centred between those of the test patterns' 1D components. Discriminability is significantly enhanced ($P < 0.05$ by Newman-Keuls tests) relative to the no-adaptation performance only for 'B' adaptors with 1D components that are parallel to those of test plaids. Systematically varying the adaptor disparity within the vicinity of the disparity of the 2D test features uncovered no region of improved discriminability; included within this region were the horizontal disparities of the test plaids' second-order gratings. Bars represent 1 s.e.m.

has long seemed oddly mismatched to the disparities of isolated objects, which have a much greater range in the horizontal than in the vertical direction. It has been suggested that the isotropy of disparity tuning serves merely to maintain ocular vergence, but it can instead be seen as an evolutionary response to the demands of stereoscopic matching in a layered (for example, arboreal) environment. □

Methods

Stimuli. In all experiments, visual patterns were constructed from achromatic sinusoidal gratings (or, in one case, squarewave gratings) with a spatial frequency of 1.0 cycle per deg and a Michelson contrast of 5 or 10%. Background luminance was 39 cd m⁻². The patterns appeared with the requisite disparity in two circular windows, one for each eye, and were viewed through a mirror stereoscope at a distance of 57 cm. A black square, 6 min on a side, was continually present at the centre of each window, as a zero-disparity fixation point. A random phase shift was added to each grating on each trial, with corresponding left- and right-eye gratings receiving the same shift, to prevent the possible learning of monocular position cues. A Macintosh computer controlled the stimulus display.

Perceived depth experiment. (See Fig. 2.) On each trial, a single pair of superimposed gratings appeared for 180 ms, with abrupt onset and offset. The contrast of each grating was 5%. Waveforms were sinusoidal in two conditions and square in a third condition. The gratings had orientations of 15° and 45° and appeared in a circular window with a diameter of 7.85° of visual angle. One of the gratings had zero disparity throughout a run of 55 trials; the other grating was given a disparity randomly selected on each trial from 11 alternative values within a phase range of $\pm \pi/4$. Observers clicked on one of two on-screen buttons, displayed 0.5 s after stimulus offset, to indicate whether the plaid (or squarewave grating) was 'near' or 'far', relative to the plane of fixation. Trials were self-paced; no feedback about accuracy of responses was given. Each data point was based on 30 trials. Three observers participated; their monocular and stereo acuities were normal or corrected-to-normal. One observer was naive as to the purposes of the experiment.

Adaptation experiment. (See Figs 4 and 5.) On each trial, two reversed-depth

stereo plaids appeared in separate 180-ms intervals separated by a 400-ms blank period. Observers selected the interval containing the more distant of the two test plaids. The plaids consisted of a 15° positive-disparity grating and a 45° zero-disparity grating, each with 5% contrast. They appeared at slightly different depths on the near side of the plane of fixation, as shown in Fig. 4. The two 15° gratings had disparity phase angles of 31.0° and 18.6°, yielding a horizontal disparity difference of 8'. The disparity of the adapting pattern was either midway between the disparities of the 15° component gratings appearing in the two test intervals or midway between the disparities of the 2D features appearing in the two test intervals. There were four adapting stimuli: a 15° grating, a 105° grating, a plaid with components at 15° and 45°, and a plaid with components at 105° and 135°. The horizontal disparities of the adaptors were set to correspond to the depths of adaptors A and B in Fig. 4. For plaid adaptors, both components (and the 2D features) had the same horizontal spatial disparity. Adapting gratings had spatial frequencies of 1.0 cycle per deg, the same as the test gratings; they appeared in a circular window with a diameter of 9°, to ensure complete retinal overlay of the 7.85° test window. Each run of 20 trials began with an adapting period of 40 s, and each trial began with a 4 s 'topping-off' adaptation period, with 0.4 s separating adaptor offset and the first test interval. Adaptors underwent smooth harmonic motion at 1.0 cycle per s to minimize retinal adaptation. Adaptation and no-adaptation conditions were identical except for the adaptor grating contrast (10 and 0%, respectively). Adaptor contrast and duration were identified by systematic sampling to favour disparity adaptation at the expense of contrast adaptation; relatively brief and low-contrast adaptors tended to yield this result. Contrast adaptation and disparity adaptation have opposite expected consequences, the former inhibiting and the latter facilitating depth discrimination. Observers were instructed to maintain fixation on the central square throughout each trial and to respond by clicking one of two on-screen buttons to indicate the interval containing the test plaid more distant from the observer. Of three observers, one was naive; all had corrected-to-normal monocular and stereo acuities. Each data point was based on 40 trials.

Received 21 November 1997; accepted 31 July 1998.

1. Wheatstone, C. Contributions to the physiology of vision. Part the first. On some remarkable, and hitherto unobserved, phenomena of binocular vision. *Phil. Trans. R. Soc. Lond. B* 2, 371–393 (1838).
2. Ogle, K. N. *Researches in Binocular Vision* (Saunders, Philadelphia, 1950).
3. Bishop, P. O. *Neurosciences Research Study Program Boulder Symposium* (ed. Schmitt, F. O.) (Rockefeller Univ. Press, New York, 1970).
4. Julesz, B. *Foundations of Cyclopean Perception* (Univ. of Chicago Press, 1971).
5. Wallach, H. *On Perception* (ed. Wallach, H.) (Quadrangle, New York, 1976).
6. Farell, B. & Ahuja, S. Binocular disparity of 1- and 2-D contours. *Invest. Ophthalmol. Vis. Sci.* (suppl.) 37, 284 (1996).
7. Farell, B. Depth coherence and depth reversal in stereo plaids. *Invest. Ophthalmol. Vis. Sci.* (suppl.) 38, 904 (1997).
8. Ramachandran, V. S. & Cavanagh, P. Subjective contours capture stereopsis. *Nature* 317, 527–531 (1985).
9. Schumier, R. A. & Ganz, L. Disparity averaging in random dot stereopsis. *J. Opt. Soc. Am.* 69, 1479 (1979).
10. Stevenson, S. B. & Schor, C. M. Human stereo matching is not restricted to epipolar lines. *Vision Res.* 37, 2717–2723 (1997).
11. Morgan, M. J. & Castet, E. The aperture problem in stereopsis. *Vision Res.* 39, 2737–2744 (1997).
12. Adelson, E. H. & Movshon, J. A. Phenomenal coherence of moving visual patterns. *Nature* 300, 523–525 (1982).
13. DeAngelis, G. C., Ohzawa, I. & Freeman, R. D. Depth is encoded in the visual cortex by a specialized receptive field structure. *Nature* 352, 156–159 (1991).
14. Blakemore, C. & Hague, B. Evidence for disparity detecting neurons in the human visual system. *J. Physiol. (Lond.)* 225, 437–455 (1972).
15. Hibbard, P. B. & Langley, K. Plaid slant and inclination thresholds can be predicted from components. *Vision Res.* 38, 1073–1084 (1998).
16. von Helmholtz, H. *Treatise on Physiological Optics* Vol. 3 (ed. Southall, J. P. C.) (translated from German 3rd edn) (Optical Society of America, 1925).
17. Gillam, B., Chambers, D. & Lawergren, B. The role of vertical disparity in the scaling of stereoscopic depth perception: an empirical and theoretical study. *Percept. Psychophys.* 44, 473–483 (1988).
18. Rogers, B. J. & Bradshaw, M. F. Vertical disparities, differential perspective and binocular stereopsis. *Nature* 361, 253–255 (1993).
19. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol. (Lond.)* 160, 106–154 (1962).
20. Barlow, H. B., Blakemore, C. & Pettigrew, J. D. The neural mechanism of binocular depth discrimination. *J. Physiol. (Lond.)* 193, 327–342 (1967).
21. Poggio, G. F. & Fischer, B. Binocular interaction and depth sensitivity in striate and prestriate cortex of behaving rhesus monkey. *J. Neurophysiol.* 40, 1392–1405 (1977).
22. Zeki, S. Functional specialization and binocular interaction in the visual areas of rhesus monkey prestriate cortex. *Proc. R. Soc. Lond. B* 204, 379–397 (1979).
23. Ferster, D. A comparison of binocular depth mechanisms in area 17 and 18 of the cat visual cortex. *J. Physiol. (Lond.)* 311, 623–655 (1981).
24. LeVay, S. & Voigt, T. Ocular dominance and disparity coding in cat visual cortex. *Vis. Neurosci.* 1, 395–414 (1988).
25. Nikara, T., Bishop, P. O. & Pettigrew, J. D. Analysis of retinal correspondence by studying receptive fields of binocular single units in cat striate cortex. *Exp. Brain Res.* 5, 353–372 (1968).

26. Anzai, A., Ohzawa, I. & Freeman, R. D. Neural mechanisms underlying binocular fusion and stereopsis: Position vs. phase. *Proc. Natl Acad. Sci. USA* 94, 5438–5443 (1997).
27. Mayhew, J. E. W. & Frisby, J. P. Stereopsis masking in humans is not orientationally tuned. *Perception* 7, 431–436 (1978).

Acknowledgements. I thank D. C. Moore for laboratory assistance, and S. Ahuja, R. van Ee, J. Krauskopf, D. Matza, D. G. Pelli and S. B. Stevenson for comments and discussion.

Correspondence and requests for materials should be addressed to B.F. (e-mail: bart_farell@isr.syr.edu).

Who reads temporal information contained across synchronized and oscillatory spike trains?

Katrina MacLeod, Alex Bäcker & Gilles Laurent

California Institute of Technology, Division of Biology 139-74, Pasadena, California 91125, USA

Our inferences about brain mechanisms underlying perception rely on whether it is possible for the brain to 'reconstruct' a stimulus from the information contained in the spike trains from many neurons^{1–5}. How the brain actually accomplishes this reconstruction remains largely unknown. Oscillatory and synchronized activities in the brain of mammals have been correlated with distinct behavioural states or the execution of complex cognitive tasks^{6–11} and are proposed to participate in the 'binding' of individual features into more complex percepts^{12–14}. But if synchronization is indeed relevant, what senses it? In insects, oscillatory synchronized activity in the early olfactory system seems to be necessary for fine odour discrimination¹⁵ and enables the encoding of information about a stimulus in spike times relative to the oscillatory 'clock'¹⁶. Here we study the decoding of these coherent oscillatory signals. We identify a population of neurons downstream from the odour-activated, synchronized neuronal assemblies. These downstream neurons show odour responses whose specificity is degraded when their inputs are desynchronized. This degradation of selectivity consists of the appearance of responses to new odours and a loss of discrimination of spike trains evoked by different odours. Such loss of information is never observed in the upstream neurons whose activity is desynchronized. These results indicate that information encoded in time across ensembles of neurons converges onto single neurons downstream in the pathway.

The function of oscillations and synchronization in information processing, perception and action is difficult to establish directly. Studies in mammals have correlated the degree of neural synchronization with specific behavioural or cognitive tasks, such as segmentation⁸, rivalry⁹, and sensorimotor tasks^{10,11}, suggesting a functional link. In locusts, stimulation by odours evokes synchronized firing in dynamic and odour-specific ensembles of projection neurons in the antennal lobe, a region analogous to the vertebrate olfactory bulb^{16–18}. This synchronization relies critically on fast GABA (γ -aminobutyric acid)-mediated inhibition, and can be selectively blocked by local injection of the GABA receptor antagonist picrotoxin¹⁹. Picrotoxin spares the slow modulation of individual projection neuron responses¹⁹ but desynchronizes the firing of the odour-activated assemblies¹⁹ and impairs fine odour discrimination¹⁵. These results establish a causal link between synchronization and perception. They do not, however, determine where information is lost when projection neurons—the information channels—are desynchronized. One possibility is that no single neuron between sensory and motor/cognitive areas is, on its own, sensitive to input synchronization. The behavioural deficit caused

by desynchronization of projection neurons induced by picrotoxin would thus be a result of collective neural activity only. Another possibility is that injection of picrotoxin leads to a loss of information in the responses of individual projection neurons due, for example, to jittering in their spike times. Desynchronization of projection neurons would thus be a by-product of picrotoxin treatment, but not the actual cause of the behavioural deficit. Finally, perhaps individual projection neuron responses undergo no loss of information after picrotoxin treatment, but responses of neurons downstream from them do. This result would indicate that information contained across projection neurons¹⁶ is indeed crucial, and it would identify the downstream neurons as decoders of this relational/temporal information.

We searched for such putative neural elements downstream from the antennal lobe projection neurons, which project to the mushroom body, where they make divergent connections onto ~50,000 mushroom body intrinsic neurons (Fig. 1a). Odours cause oscillatory activity in these neurons, but they evoke spiking responses only in very sparsely distributed ones²⁰, making the sampling of odour-responsive neurons very difficult. We thus focused on a smaller population of neurons directly postsynaptic to the intrinsic neurons of the mushroom body^{21,22}, two synapses downstream from the projection neurons (Fig. 1b). These neurons, the β -lobe neurons, receive convergent input from thousands of mushroom body intrinsic neurons and have clear odour-specific responses (Fig.

1c–e). This makes them a suitable ‘read-out’ of signals processed in the early olfactory system. We recorded intracellularly from the dendrites of these β -lobe neurons. Dye injection showed discrete and dense (presumed dendritic) arbors in the β -lobe and sometimes also within the pedunculus (Fig. 1b, left). Sparser, varicose (presumed axonal) fibres projected to the α -lobe and sometimes also within the pedunculus (Fig. 1b, left). Many morphological types were found, indicating a heterogeneous population. Morphologically identical examples of a type (for example, Fig. 1b, left) could have different physiological profiles in different animals, indicating either many exemplars of a type in each animal, or an animal-specific tuning history. Many, though not all, β -lobe neurons responded to at least some of the ten odours that we presented and about half of them showed significant phase-locking of their spikes to the odour-evoked local field potential (LFP) oscillations. Responses were usually stimulus-specific but less complex than those of the projection neurons of the antennal lobe¹⁸. They consisted of phasic or phasotonic increases in firing rate and, in a few cases, of a suppression of firing (Fig. 1c–f). Whereas spiking responses could be brief, a subthreshold synaptic drive consisting of both excitatory and inhibitory postsynaptic potentials (EPSPs and IPSPs) was usually maintained throughout the duration of the stimulus and mushroom body LFP oscillation (Fig. 1e).

To determine whether responses of the β -lobe neurons depend on synchronization of the projection neurons, a β -lobe neuron was

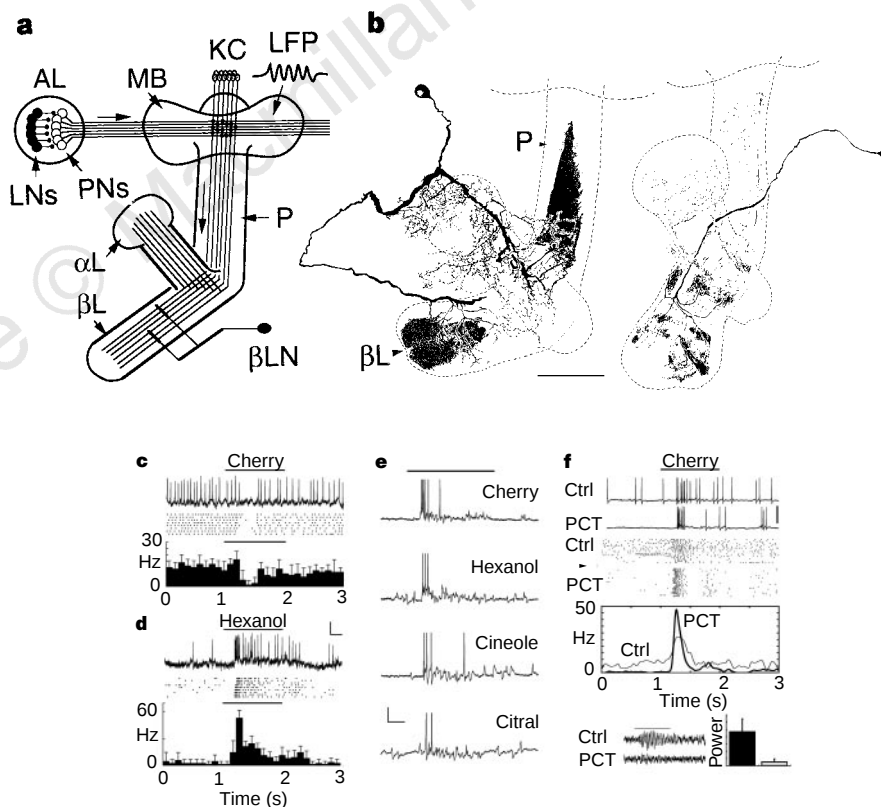


Figure 1 β -lobe neurons (β LNs). **a**, The insect olfactory pathways. Antennal lobe (AL) projection neurons (PNs) receive excitatory input from peripheral olfactory receptor neurons and GABA-mediated inhibitory input from local neurons (LNs)^{19,25}. PNs project to the mushroom body (MB) calyx, where oscillatory local field potentials (LFPs) can be recorded in response to odours. Intrinsic neurons of the MB (the Kenyon cells, KCs) receive direct excitatory input from PNs^{20,25}, and bifurcate to the α - and β -lobes (α L, β L)²⁰. β LNs receive input from thousands of KCs. **b**, Left and right, cobalt hexamine fills of two β LNs, each from a different animal. P, pedunculus of the MB. Calibration, 100 μ m. **c**, Suppression of activity of

the left β LN in **b** in response to a cherry odour. Top, intracellular; centre, rasters; bottom, PSTH; mean firing $\pm$ s.d. All odour pulses lasted for 1 s; 7–10 s between trials. Calibrations, 5 mV, 0.2 s. **d**, Prolonged excitatory response of the right β LN in **b** in response to the odour hexanol. Calibrations, 2 mV, 0.2 s. **e**, Transient excitatory responses of a third β LN to four odours. Note subthreshold activity during stimulus. Calibrations, 10 mV, 0.2 s. **f**, Effect of picrotoxin injection on a fourth β LN. Top to bottom: intracellular, rasters, smoothed PSTHs, LFP (bottom left), and normalized integrated power of LFP (bottom right) over 15–30 Hz, mean $\pm$ s.d., before (black) and after (open) picrotoxin injection. Calibration, 50 mV.

impaled and its odour tuning characterized. Picrotoxin was injected into the ipsilateral antennal lobe and the induced PN desynchronization was verified from the power spectrum of the ipsilateral mushroom body LFPs¹⁹. The responses of the β -lobe neurons to odours were then tested again. We made three main observations ($n = 19$ β -lobe neurons). First, picrotoxin-induced desynchronization of projection neurons never abolished those β -lobe neuron odour responses that existed before injection of picrotoxin ($n = 15$ β -lobe neurons). Second, desynchronization of projection neurons often caused changes in some, though not all existing β -lobe neuron odour responses. These changes consisted of a decrease in baseline firing rates ($n = 6$ β -lobe neurons), or a combined increase in the response and a decrease in the spontaneous firing rates ($n = 4$ β -lobe neurons), resulting in a more 'sculpted' response than in pre-picrotoxin controls (Fig. 1f). In many cases, injection of picrotoxin caused a change in the temporal patterning of the β -lobe neuron odour response ($n = 14$ β -lobe neurons) (see below). Third, desynchronization of the projection neurons could induce a β -lobe neuron to respond to odours to which it was previously not tuned (Fig. 2). Such picrotoxin-induced changes occurred in five β -lobe neurons and in nine odour–neuron combinations out of the fourteen tested; for example, a β -lobe neuron could, after desynchronization of the projection neurons, acquire a response to one odour and not to another. This selectivity was observed in four of the five β -lobe neurons. Such picrotoxin-induced responses were never observed in antennal lobe projection neurons ($n = 10$ projection neurons). In summary, some changes induced in β -lobe neurons by injection of picrotoxin into the antennal lobe consisted of the appearance of *de novo* odour sensitivity, thus decreasing the

neuron's selectivity. Other changes (in 58 per cent of all neurons and 34 per cent of all odour–neuron pairs tested) were more subtle and consisted of a modification of existing responses to a set of odours (see below).

Consider a β -lobe neuron that shows different response patterns to each of three odours, A, B and C. Desynchronization of the projection neurons often caused the responses to A, B, and C (or to a pair of these odours) to become more similar to each other, thus decreasing the likelihood of correct odour identification using the information contained in each spike train. We used two clustering algorithms to quantify the effect of picrotoxin treatment on β -lobe neuron response classification. These algorithms categorize trials and assign each one to an odour on the basis of similarity between spike trains (see Methods). An example is shown in Fig. 3a, in which desynchronization of projection neurons caused a considerable loss in the specificity of the β -lobe neuron response to three related aliphatic alcohols. This classification (Fig. 3b) was carried out with pairwise comparisons of all pre-picrotoxin trials (Fig. 3c, top row) and indicated a complete discriminability (Fig. 4a, left). Picrotoxin was injected into the ipsilateral antennal lobe and classification was repeated; that is, each response was attributed a probable stimulus given its proximity to any of the three pre-picrotoxin templates. The

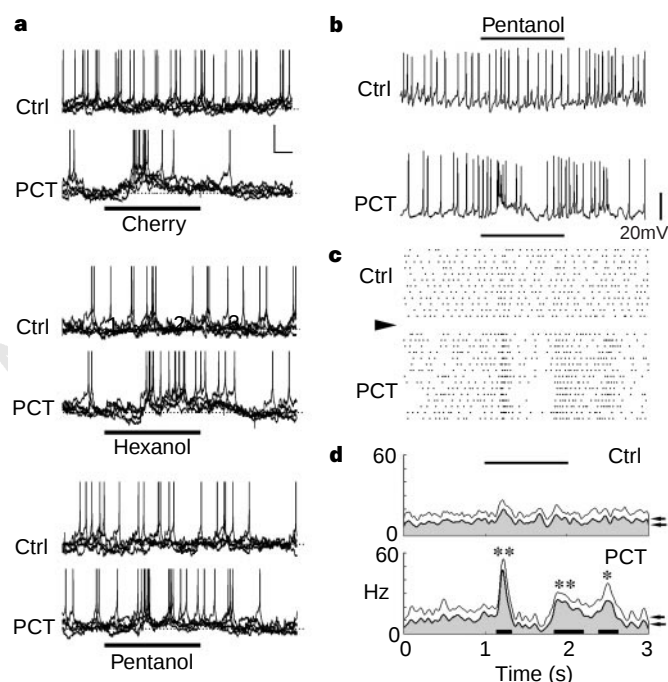


Figure 2 Induction of odour-evoked responses in β -lobe neurons (BLNs) after injection of picrotoxin (PCT) into the antennal lobe. **a**, Five superimposed intracellular traces from one BLN before (Ctrl) and after PCT injection, for three different odours. Dotted line, resting membrane voltage. Calibrations: 5 mV, 0.2 s. **b–d**, Measurements from a different BLN. **b**, Intracellular traces. **c**, Rasters. **d**, Smoothed PSTH. Although there was no response to pentanol in the control, this BLN developed a response to this odour after PCT injection. Shaded PSTH represents mean rate; thin line represents s.d. Dark bars at bottom represent epochs in which the firing rate is significantly different from control. $**P < 0.01$; $*P < 0.05$.

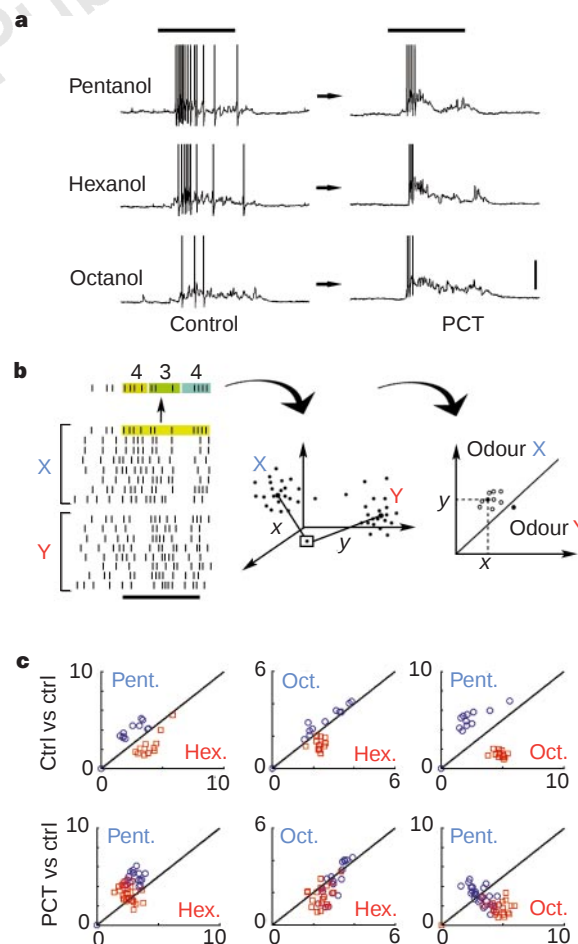


Figure 3 Desynchronization of projection neurons causes loss of information in odour-evoked spike trains in single β -lobe neurons (BLN). **a**, Intracellular traces from one BLN in response to three odours before (control) and after picrotoxin (PCT) treatment. Note the decrease in the difference between the three spike trains after PCT treatment. **b**, The clustering algorithm (method 1, Methods). **c**, Clustering algorithm (20-dimensional space) applied to control spike trains (top row) or post-PCT spike trains (bottom row) for the BLN in **a**. Diagonal line delineates classification, with results shown in Fig. 4a.

pairwise (and overall) classification now operated just above chance level (Figs 3c, bottom row and 4a, left). This indicates that picrotoxin induced both a shift in odour representation by the β -lobe neurons and a loss of information about stimulus identity in the temporal features of their spike trains. Among all β -lobe neurons for which there was good odour discriminability in controls ($n = 14$ β -lobe neurons; $n = 34$ odour pairs), injection of picrotoxin hindered discrimination in eight β -lobe neurons for at least one odour pair in each neuron and for a total of nineteen odour pairs. In five of these eight neurons, discrimination was reduced to chance. Among all β -lobe neurons the probability of correct pairwise odour discrimination dropped from an average of $\sim 90\%$ to $\sim 75\%$ (Fig. 4b, left). A degradation was also found using

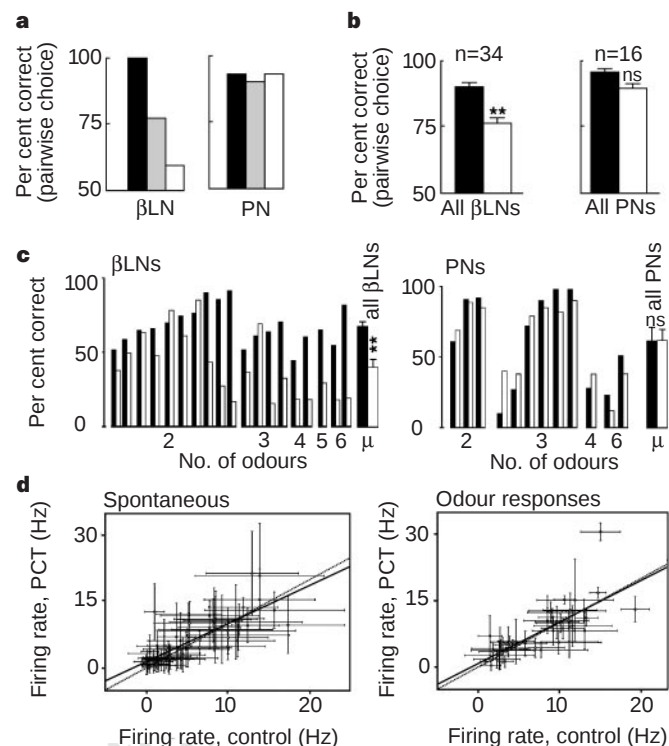


Figure 4 Odour classification by β -lobe neurons (BLNs) and projection neurons (PNs) before and after picrotoxin (PCT) treatment. **a**, Categorization over all trials for the BLN in Fig. 3 and the PN in Fig. 5, using method 1. Black, control vs control; grey, PCT vs PCT; open, PCT vs control. Left, Perfect categorization of control trials (black, 72 correct out of 72) but impaired classification after PCT injection (grey, 147 of 190 trials; open, 113 of 190 trials). Right, PN shows an equally good categorization under all conditions (black, 112 of 120; grey, 121 of 134; open, 124 of 134). **b**, Summary of pairwise categorizations (method 1) for all BLNs (left) and PNs (right) with significant categorization of control trials. Black, control vs control; open, PCT vs control. Left, BLNs show a significant decrease in the percentage of correct categorization after PCT injection ($**P < 0.005$). Right, PN performance is not significantly altered by PCT treatment. **c**, Categorization results using clustering method 2, applied to all BLNs (left) and all PNs (right) with significant categorization of control trials. 'Per cent correct' was calculated using all possible choices. BLNs and PNs are ranked on the abscissa according to the number of odours (2–6) to which they responded. μ is the mean over all BLNs (left, $**P < 0.0002$) and all PNs (right, $P = 0.93$, ns). Each neuron and μ are represented by a pair of bars: black, control vs control; open, PCT vs control. **d**, PCT evokes no overall change in PN baseline (left, $n = 54$ odour-neuron pairs) or odour-evoked (right, $n = 38$ odour-neuron pairs) mean firing rates. Each point plots rate in one PN before (x-axis) vs after (y-axis) PCT injection. Rates (mean $\pm$ s.d.) calculated over 1 s, from 8–30 trials from 26 PNs. We included only examples in which odour caused a response. Slope of linear regression fits: 0.86 (left, $R = 0.8$); 0.94 (right, $R = 0.76$).

post-picrotoxin response patterns as templates (for example, Fig. 4a, left). Such discrimination was impaired (decreased by 10% or more) in seven β -lobe neurons ($n = 9$ odour pairs) and reduced to chance in three.

To determine whether the information loss induced by picrotoxin in β -lobe neurons can be attributed to information loss in individual projection neurons, the same analysis was carried out with the projection neurons (Fig. 5a), for example, responded to three odours with three distinct temporal patterns, allowing 93% correct odour identification in paired choices (Figs 4a, right and 5b, top row). After injection of picrotoxin into the antennal lobe, this projection neuron retained its original odour-specific response patterns, allowing a correct odour identification using either pre- or post-picrotoxin templates (Figs 4a, right and 5b, bottom row). In all the projection neurons thus tested, that showed above-chance odour discrimination before picrotoxin injection (16 odour/projection neuron pairs), injection of picrotoxin failed to alter pairwise odour discrimination using information contained in the projection neuron spike trains (Fig. 4b, right). A further comparison of odour classification by projection and β -lobe neurons was carried out using a clustering algorithm that preserves the continuous structure of the spike trains (see Methods). Over all β -lobe neurons ($n = 19$) and odours (2.6 per neuron on average) tested, correct classification using this method dropped significantly from 67.6% to 40.1% (chance = 38.5%) after desynchronization of projection neurons (Fig. 4c, left). Over all projection neurons ($n = 12$) and odours (2.9 on average per neuron) tested, odour

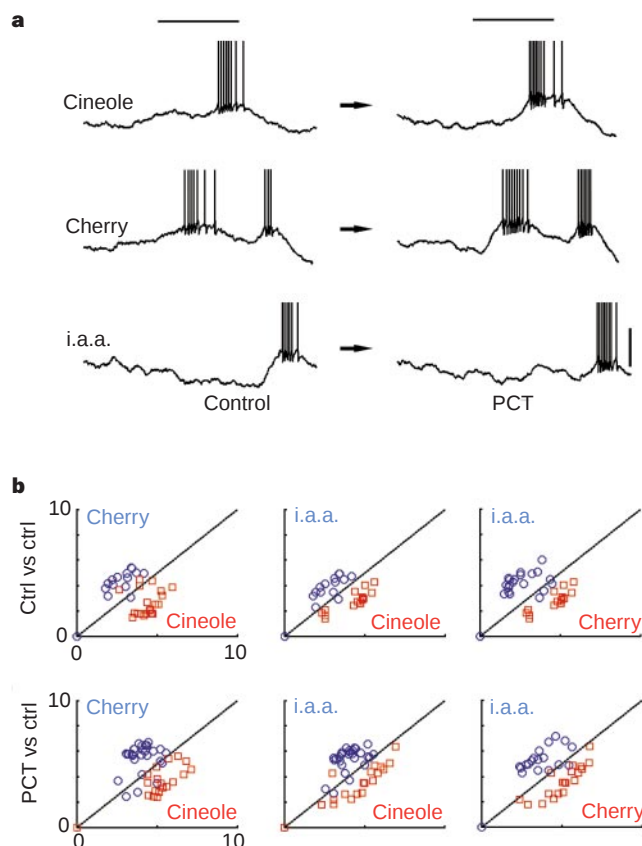


Figure 5 Desynchronization of projection neurons (PNs) causes no loss of information in odour-evoked spike trains in single PNs. The same analysis as that in Fig. 3 was performed, but on a single PN. **a**, Note distinct response patterns to different odours (i.a.a., isoamylacetate) and their persistence after picrotoxin (PCT) treatment. **b**, Clustering algorithm (20-dimensional space) applied to control spike trains (top row) or post-PCT spike trains (bottom row) for the PN in **a**. Classification results are shown in Fig. 4a.

classification was unaffected after desynchronization of projection neurons (control, 61.8% correct; picrotoxin, 62.1% correct) (Fig. 4c, right). Finally, we quantified the effects of picrotoxin injection on the spontaneous and stimulus-evoked firing rates of 26 projection neurons. These results indicated unchanged projection neuron rates in post-picrotoxin trials (Fig. 4d).

Thus, we have identified a population of central olfactory neurons that can be used as a neural 'read-out' of population activity in the antennal lobe. The tuning of these β -lobe neurons was affected by desynchronization of projection neurons induced by picrotoxin in two principal ways, which were never observed in individual projection neurons. First, odour discrimination using information contained in the spike trains of β -lobe neurons became less likely (often reduced to chance) after picrotoxin treatment. These effects were odour-specific and could affect the responses to related chemical stimuli, blurring their specific differences. Second, picrotoxin-induced desynchronization of projection neurons could cause the induction of new, but not generalized, odour responses in β -lobe neurons. Because these effects were never observed in projection neurons, we propose that they were induced by desynchronization of projection neurons and that β -lobe neurons (and perhaps the neurons presynaptic to them) are sensitive to the temporal structure of their inputs. Our earlier behavioural study indicated that treatment with picrotoxin impairs fine odour discrimination¹⁵. Our present results indicate that behavioural impairments may be due to a partial loss of stimulus-related information in neurons involved in processing information between early and higher olfactory centers. We thus propose that neural synchronization is indeed important for information processing in the brain: it is, at least partly, a temporal substrate for the transmission of information that is contained across co-activated neurons (relational code). Such information cannot be retrieved by simple averaging of the input channels' activities without considering their temporal relationships. It will be interesting to determine whether such coding principles apply to other systems endowed with neuronal synchronization. $\square$

Methods

Electrophysiology. Sixty-three adult non-anaesthetized male locusts (*Schistocerca americana*) were taken from a crowded colony and immobilized with one antenna intact and exposed for stimulation, as described¹⁸. Intracellular recordings were made from the soma or dendrites of 26 antennal lobe projection neurons and from the dendrites of 37 β -lobe neurons by using glass micropipettes¹⁸. Intracellular staining was carried out by iontophoretic injection of 6% cobalt hexamine (0.5–2 nA, 1 Hz). Histology and intensification were performed as in ref. 23. β -lobe neurons may be homologous to multimodal neurons described in other insect species^{21,22}. Local field potentials were recorded with saline-filled patch pipettes from the mushroom body calyx. Picrotoxin was applied locally to the antennal lobe neuropil by pressure injection¹⁹. Glass capillaries were back-filled with a 0.1 M saline solution of picrotoxin and injected volumes were ~ 1 pl. Controls for absence of picrotoxin leakage outside the antennal lobe were carried out by dye injection. Controls for absence of mechanical effects of drug injection were done by testing eight β -lobe neurons before and after injection of saline in the antennal lobe. In a few experiments with projection neurons, picrotoxin was bath-applied, which was shown previously to have the same effects on them as pressure-injection¹⁹. Bath application of picrotoxin was never used with β -lobe neurons. Electrophysiological data were digitized at 5 kHz post-hoc from DAT with LabVIEW software and an NBMIO16L interface (National Instruments). Data were analysed with MATLAB (Mathworks). Airborne odours were delivered to an antenna as described¹⁸.

Test for induced responses. For each set of trials (odour, condition) we calculated a mean $\pm$ s.e.m. smooth peristimulus time histogram (PSTH) by convolving each raster with a 25-ms Gaussian and summing the results. We then calculated, for each time t (1-ms resolution), whether the firing rate during the response (for 2 s after odour onset) was different from the mean baseline firing rate (for 1 s before odour onset) by at least 2 s.e.m. for at least 100

contiguous ms. The response was described as different between pre- and post-picrotoxin injection if a significant ($P < 0.05$, unpaired two-tailed t -test) difference between corresponding epochs occurred. Changes so measured could thus occur over multiple epochs within a response (Fig. 2d). This method ensured the detection of brief responses or ones that include both excitation and inhibition.

Cluster analysis, method 1. Methods 1 and 2 were used with all β -lobe and projection neurons that responded to two or more of the ten odours presented. With method 1, odour discriminations were tested pairwise (for example, four odours meant six pairwise comparisons). Each spike train in a set (one set for each pre- and post-picrotoxin condition and odour condition) was divided into n non-overlapping bins over the 2-s post-stimulus onset period. The spike count in each bin was used to construct an n -dimensional vector representing this trial. The operation was repeated with each trial in the set, to form a cluster. We tested bin sizes between 10 and 1,800 ms, and found optimal clustering in most experiments for 50–200 ms and least clustering below 10 ms or above 1,200 ms (approximately average firing rate). We present data for 100-ms bins, and thus $n = 20$. The objective was to assess the separation between two clusters (for example, between the X and Y clusters), by assigning each spike train (or vector) one likely stimulus on the basis of its Euclidian distance to the centre of each of the two clusters. For each pairwise comparison, therefore, each trial produced two numbers (for example, distances x and y to the centres of the X and Y clusters), which were plotted against each other (Fig. 3b). The per cent of correct classification over all trials was given by the proportion of points that fell on the 'correct' side of the diagonal. Significance of the 'per cent correct classification' was calculated by finding its binomial probability, given a forced choice (threshold $P < 0.05$; paired one-tailed t -tests). Control trials were compared to control (1) and to picrotoxin (2) trials (Figs 3c and 5b). Picrotoxin trials were compared to picrotoxin trials (3) (not shown). Comparison (1) describes the classification of odour responses in control conditions. Comparison (2) describes the classification of post-picrotoxin odour responses based on their control templates. This comparison best assesses the information loss caused by input desynchronization. Comparison (3) describes the classification of post-picrotoxin odour responses based on their post-picrotoxin templates.

Cluster analysis, method 2. This clustering analysis is based on the cost-based metric methods²⁴ according to which a 'distance' is computed between spike trains. This distance is defined as the cost paid to transform one spike train into the other using three elementary steps: insertion; deletion of a spike (each at a cost of 1); and displacement of a spike by 1 ms (cost of $2/T$ for each displacement, where T is the maximum separation in ms allowed between the spike time in one train and that in the other). We used a range for T between 16 and 4,000, with $T = 150$ providing the best classification overall, as with method 1. Results were not greatly different for $16 < T < 1,000$. Classification was similar whether the mean was arithmetic (all points equal) or geometric (less weight to outliers). Per cent correct in Fig. 4c results from choosing among all odours without restriction to pairwise assignments in those neurons that responded to more than two odours. For each neuron i , chance level is thus $1/m_i$, where m_i is the number of odours to which neuron i responded. The effective number of odours in the plots of μ (Fig. 4c) is calculated as $(\langle 1/m_i \rangle)^{-1}$. We used paired two-tailed t -tests to measure significance.

Received 5 June; accepted 10 August 1998.

- Mountcastle, V. B. Modality and topographic properties of single neurons of cat's somatic sensory cortex. *J. Neurophysiol.* **20**, 408–434 (1957).
- Sparks, D. L., Lee, C. & Rohrer, W. H. Population coding of the direction, amplitude and velocity of saccadic eye movements by neurons in the superior colliculus. *Cold Spring Harb. Symp. Quant. Biol.* **55**, 805–811 (1990).
- Georgopoulos, A. P., Lurito, J. T., Petrides, M., Schwartz, A. B. & Massey, J. T. Mental rotation of the neuronal population vector. *Science* **243**, 234–236 (1989).
- Newsome, W. T., Britten, K. H. & Movshon, J. A. Neuronal correlates of a perceptual decision. *Nature* **341**, 52–54 (1989).
- Abbott, L. F. Decoding neuronal firing and modeling neural networks. *Q. Rev. Biophys.* **27**, 291–331 (1994).
- Abeles, M., Bergman, H., Margalit, E. & Vaadia, E. Spatiotemporal firing patterns in the frontal cortex of behaving monkeys. *J. Neurophysiol.* **70**, 1629–1638 (1993).
- Gray, C. M. & Singer, W. Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proc. Natl Acad. Sci. USA* **86**, 1698–1702 (1989).
- Roelfsema, P., Engel, A. K., König, P. & Singer, W. The role of neuronal synchronization in response selection: a biologically plausible theory of structural representations in the visual cortex. *J. Cog. Neurosci.* **8**, 603–625 (1996).
- Fries, P., Roelfsema, P., Engel, A. K., König, P. & Singer, W. Synchronization of oscillatory responses in visual cortex correlates with perception in interocular rivalry. *Proc. Natl Acad. Sci. USA* **94**, 12699–12704 (1997).

10. Riehle, A., Grün, S., Diesmann, M. & Aertsen, A. Spike synchronization and rate modulation differentially involved in motor cortical function. *Science* **278**, 1950–1953 (1997).
11. Nicolelis, M. A. L., Baccala, L. A., Lin, R. C. S. & Chapin, J. K. Sensorimotor encoding by synchronous neural ensemble activity at multiple levels of the somatosensory system. *Science* **268**, 1353–1358 (1995).
12. Milner, P. A model for visual shape recognition. *Psychol. Rev.* **81**, 521–535 (1974).
13. von der Malsburg, C. *The Correlation Theory of Brain Function* (Internal report, Max-Planck-Institut. Biophys. Chem., Göttingen, 1981).
14. Singer, W. Synchronization of cortical activity and its putative role in information processing and learning. *Annu. Rev. Physiol.* **55**, 349–374 (1993).
15. Stopfer, M., Bhagavan, S., Smith, B. H. & Laurent, G. Impaired odour discrimination on desynchronization of odour-encoding assemblies. *Nature* **390**, 70–74 (1997).
16. Wehr, M. & Laurent, G. Odour encoding by temporal sequences of firing in oscillating neural assemblies. *Nature* **384**, 162–166 (1996).
17. Laurent, G. & Davidowitz, H. Encoding of olfactory information with oscillating neural assemblies. *Science* **265**, 1872–1875 (1994).
18. Laurent, G., Wehr, M. & Davidowitz, H. Temporal representation of odours in an olfactory network. *J. Neurosci.* **16**, 3837–3847 (1996).
19. MacLeod, K. & Laurent, G. Distinct mechanisms for synchronization and temporal patterning of odour-encoding neural assemblies. *Science* **274**, 976–979 (1996).
20. Laurent, G. & Naraghi, M. Odorant-induced oscillations in the mushroom bodies of the locust. *J. Neurosci.* **14**, 2993–3004 (1994).
21. Homberg, U. Processing of antennal information in extrinsic mushroom body neurons of bee brain. *J. Comp. Physiol. A* **154**, 825–836 (1984).
22. Li, Y. & Strausfeld, N. Morphology and sensory modality of mushroom body extrinsic neurons in the brain of the cockroach, *Periplaneta americana*. *J. Comp. Neurol.* **387**, 631–650 (1997).
23. Bacon, J. P. & Altman, J. S. A silver intensification method for cobalt-filled neurones in wholemount preparations. *Brain Res.* **138**, 359–363 (1977).
24. Victor, J. D. & Purpura, K. P. Metric-space analysis of spike trains: theory, algorithms and application. *Network Comp. Neural. Syst.* **8**, 127–164 (1997).
25. Leitch, B. & Laurent, G. GABAergic synapses in the antennal lobe and mushroom body of the locust olfactory system. *J. Comp. Neurol.* **373**, 487–514 (1996).

Acknowledgements. We thank E. M. Schuman and C. Koch for comments on the manuscript, and the members of the Laurent laboratory for constructive criticism during this work. This work was supported by NSF and NIDCD grants (to G.L.), the Sloan Center for Neuroscience at Caltech and an NIH training grant.

Correspondence and requests for materials should be addressed to G.L. (e-mail: laurentg@cco.caltech.edu).

Destabilization of β -catenin by mutations in presenilin-1 potentiates neuronal apoptosis

Zhuohua Zhang^{*}, Henrike Hartmann^{*}, Viet Minh Do^{*}, Dorothee Abramowski[†], Christine Sturchler-Pierrat[†], Matthias Staufenbiel[†], Bernd Sommer[†], Marc van de Wetering[‡], Hans Clevers[‡], Paul Saftig[§], Bart De Strooper^{||}, Xi He^{*} & Bruce A. Yankner^{*}

^{*} Department of Neurology, Harvard Medical School and Division of Neuroscience, The Children's Hospital, 300 Longwood Avenue, Boston, Massachusetts 02115, USA

[†] Preclinical Research, Novartis Pharma Ltd, CH-4002 Basel, Switzerland

[‡] Department of Immunology, University Hospital, 3508 GA Utrecht, The Netherlands

[§] Zentrum Biochemie und Molekular Zellbiologie, Abteilung Biochemie II, Universität Göttingen, 37073 Göttingen, Germany

^{||} Flemish Institute for Biotechnology (VIB4), Center for Human Genetics, K.U. Leuven, B-3000 Belgium

Mutations of the presenilin-1 gene are a major cause of familial early-onset Alzheimer's disease^{1–4}. Presenilin-1 can associate with members of the catenin family of signalling proteins, but the significance of this association is unknown^{5,6}. Here we show that presenilin-1 forms a complex with β -catenin *in vivo* that increases β -catenin stability. Pathogenic mutations in the presenilin-1 gene reduce the ability of presenilin-1 to stabilize β -catenin, and lead to increased degradation of β -catenin in the brains of transgenic mice. Moreover, β -catenin levels are markedly reduced in the brains of Alzheimer's disease patients with presenilin-1 mutations. Loss of β -catenin signalling increases neuronal vulnerability to apoptosis induced by amyloid- β protein. Thus, mutations in presenilin-1 may increase neuronal apoptosis by altering the

stability of β -catenin, predisposing individuals to early-onset Alzheimer's disease.

To determine whether presenilin-1 (PS1) interacts with β -catenin in a physiological complex, we studied immunoprecipitates of endogenous PS1 from human cell lines and brain for the presence of β -catenin. Antibodies against both the amino- and the carboxy-terminal domains of PS1 (PS1-N and PS1-C)⁷ precipitated both PS1 and β -catenin in every cell line studied and in human brain (Fig. 1a, b). We confirmed the specificity of immunoprecipitation of PS1 by pre-absorption of the PS1-N and PS1-C antibodies with their cognate antigens (Fig. 1b), and by immunoprecipitations with PS1 preimmune serum or with an antibody to SREBP-1, which is, like PS1, a transmembrane protein of the endoplasmic reticulum (Fig. 1d).

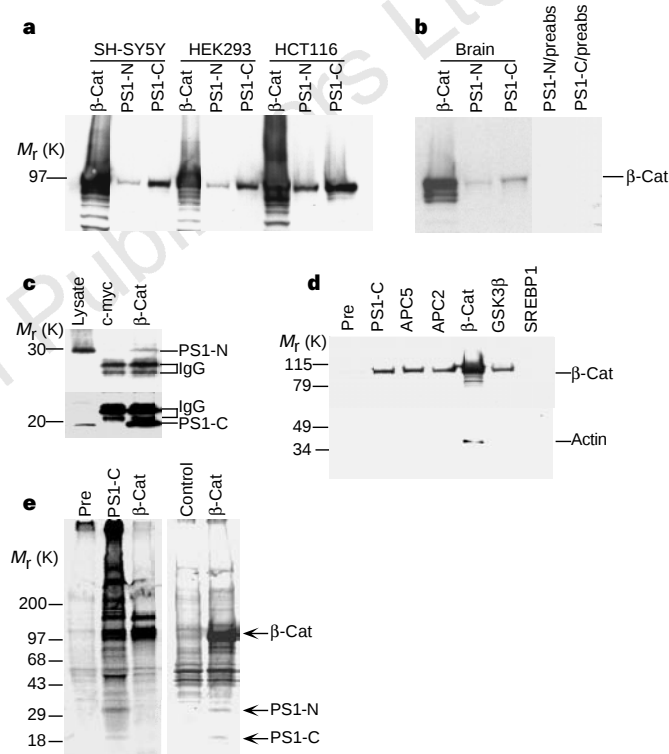


Figure 1 Native complex formation between β -catenin and PS1. **a**, Co-precipitation of β -catenin with PS1 in neuroblastoma SH-SY5Y, HEK293 and HCT116 cells. Protein-equivalent cell lysates were immunoprecipitated with antibodies against β -catenin (β -Cat), the PS1 N terminus (PS1-N), or the PS1 C-terminal loop domain (PS1-C), and then analysed for β -catenin by immunoblotting with an anti- β -catenin monoclonal antibody (mAb14). The immunoprecipitating antibodies are indicated above each lane and the immunoblotting antibodies are indicated at the right of **b**. **b**, Co-precipitation of β -catenin with PS1 in brain. Pre-absorption of the anti-PS1 antibodies by their cognate antigens abolishes co-precipitation of β -catenin (PS1-N/preabs, PS1-C/preabs). **c**, Co-precipitation of the PS1 N- and C-terminal fragments with β -catenin. HCT116 cell lysates were immunoprecipitated with a control antibody (c-Myc) or with anti- β -catenin (β -Cat), followed by immunoblotting with anti-PS1-N (upper panel) or anti-PS1-C (lower panel). Immunoblotting of the whole-cell lysate shows the positions of the PS1 N- and C-terminal fragments (PS1-N and PS1-C; upper and lower panels, respectively). The IgG light chain from the immunoprecipitation is indicated. **d**, Comparative co-precipitation of PS1, β -catenin, GSK-3 β and APC. Immunoprecipitations of protein-equivalent lysates of SW480 cells were done with preimmune PS1 serum (Pre), or with the indicated antibodies, followed by immunoblot analysis for β -catenin or actin. **e**, Stoichiometric relation of PS1 to β -catenin in PS1- β -catenin complexes. COS cells co-transfected with β -catenin and PS1 (left panel) or non-transfected SW480 cells (right panel) were metabolically labelled with [³⁵S]methionine and immunoprecipitated with the indicated antibodies. The co-precipitated PS1 N- and C-terminal fragments and β -catenin are shown. A stoichiometry of 2:1 (PS1-C: β -catenin) was determined for both cell types.

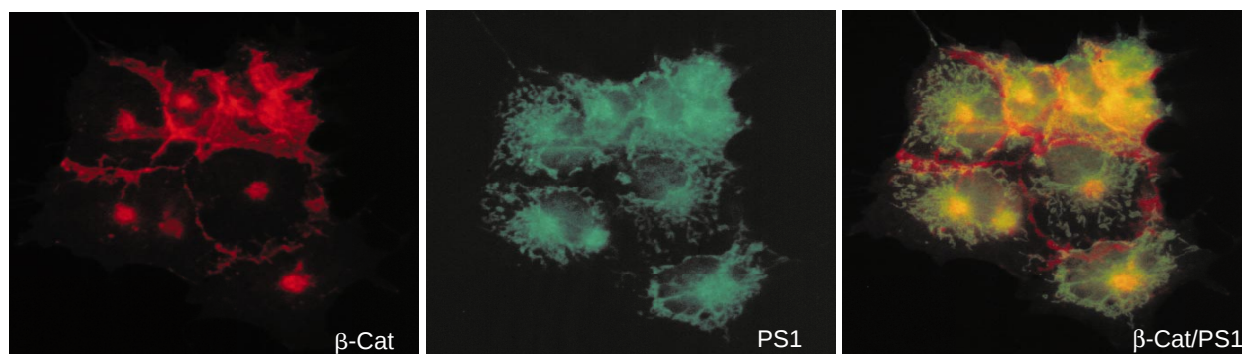


Figure 2 Co-localization of PS1 and β -catenin in the Golgi complex. COS cells were double-labelled with anti- β -catenin (MAB14; red) and anti-PS1-C (green) antibodies. Endogenous β -catenin is mainly localized to the plasma membrane

and Golgi complex (top panel), whereas endogenous PS1 is localized to the endoplasmic reticulum and Golgi (middle panel). β -Catenin and PS1 co-localize in the Golgi complex (yellow in bottom panel).

The reversed protocol, immunoprecipitation of β -catenin followed by immunoblot analysis for PS1, demonstrated co-precipitation of β -catenin with the N- and C-terminal fragments of PS1 (ref. 8), but not with the full-length form of PS1 (Fig. 1c). These results indicate that β -catenin is present in one or more complexes with the N- and C-terminal fragments of PS1.

We next compared the levels of β -catenin–PS1 complexes with those of complexes containing β -catenin and glycogen synthase kinase-3 β (GSK-3 β) or adenomatous polyposis coli protein (APC), as GSK-3 β and APC form physiological complexes with β -catenin that are involved in signal transduction⁹. The amount of β -catenin that co-precipitated with PS1 was similar to the amount that co-precipitated with GSK-3 β and APC (Fig. 1d), indicating that PS1

may be associated with physiologically relevant levels of β -catenin. The stoichiometry of the PS1– β -catenin complex was 2:1 (PS1-C: β -catenin), as determined by metabolic labelling with [³⁵S]methionine followed by co-precipitation of PS1 with β -catenin (Fig. 1e).

We then identified the subcellular localization of the β -catenin–PS1 complex. Immunofluorescence microscopy of endogenous PS1 showed that it was mainly distributed in the endoplasmic reticulum and Golgi (Fig. 2), as described previously^{7,10,11}. β -Catenin was localized predominantly to the cell surface and Golgi complex (Fig. 2). Double-labelling showed that PS1 co-localized with β -catenin in the Golgi complex. Thus, PS1 may sequester cytoplasmic β -catenin in a membrane-associated complex.

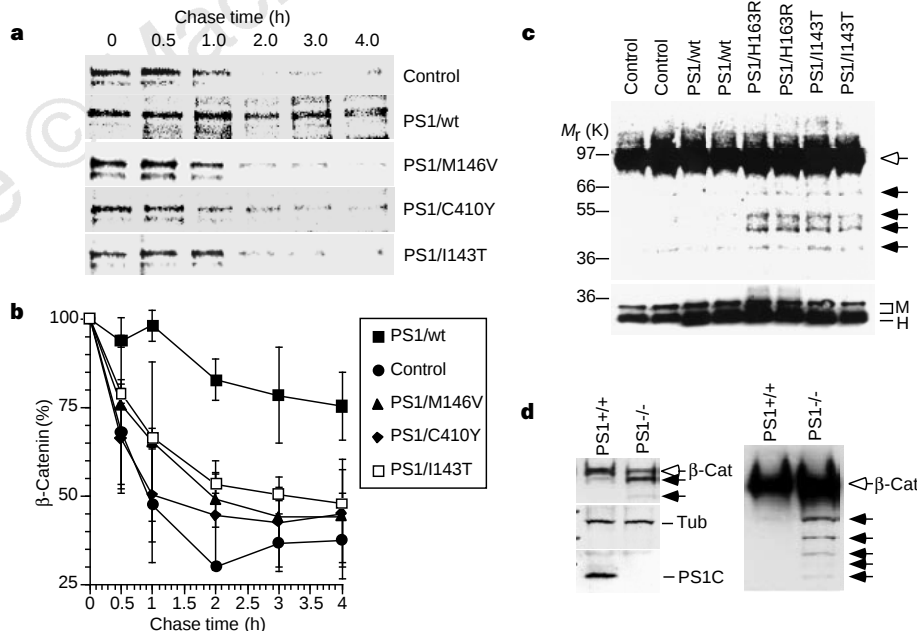


Figure 3 Stabilization of β -catenin by PS1. **a**, Pulse-chase analysis of β -catenin degradation. HEK293 cells were co-transfected with Myc-tagged β -catenin and either wild-type (wt) PS1, mutant PS1 (PS1/M146V, PS1/C410Y or PS1/I143T) or the empty vector cDNA (control); transfection was followed by pulse-chase labelling with [³⁵S]methionine and analysis by immunoprecipitation of β -catenin. **b**, Quantitative analysis of β -catenin pulse-chase kinetics. Values are the per cent of the amount of β -catenin at time 0 (100%) and represent the mean $\pm$ s.e.m. Similar results were obtained in four independent experiments. **c**, Increased β -catenin-degradation products in the brains of mice transgenic for PS1 mutations. Upper panel, immunoblot analysis of β -catenin in frontal cortex. Full-length β -catenin (open arrow) and β -catenin-degradation products (filled arrows) are

indicated. Results shown are from two lines of mice transgenic for wild-type PS1 or the PS1 mutations H163R and I143T, and from non-transgenic littermate controls. Lower panel, the same immunoblot analysed with anti-PS1-N antibody, showing the PS1 N-terminal fragments. The endogenous mouse (M) and transgenic human (H)-derived N-terminal fragments are indicated. **d**, Destabilization of β -catenin in PS1-deficient fibroblasts. Left, immunoblot analysis of β -catenin in embryonic fibroblast cultures from PS1-null (PS1^{-/-}) and control (PS1^{+/+}) mice. The level of full-length β -catenin (open arrow) is decreased in PS1-deficient cells, whereas the level of tubulin (Tub) is unaffected. Right, longer exposure reveals a ladder of β -catenin-degradation products (filled arrows) in PS1-deficient cells that is not detected in wild-type cells.

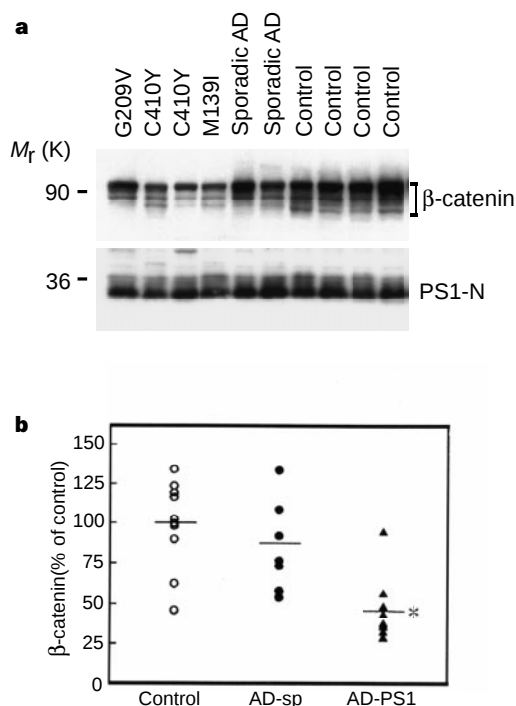


Figure 4 Reduced levels of β -catenin in the brains of Alzheimer's disease patients with PS1 mutations. **a**, Immunoblot analysis of β -catenin (upper panel) and PS1 (lower panel) in the temporal cortex of Alzheimer's disease patients with the indicated PS1 mutations, two patients with sporadic Alzheimer's disease, and four controls without Alzheimer's disease. **b**, Quantification of β -catenin in temporal cortex. Values are percentages of the control mean of β -catenin (control, $n = 10$; sporadic Alzheimer's disease (AD-sp), $n = 7$; Alzheimer's disease with PS1 mutations (AD-PS1), $n = 9$). Asterisk indicates $P < 0.001$ relative to control by ANOVA with post-hoc Student–Newman–Keuls test. Linear regression analysis indicates the absence of any correlation between β -catenin level and age or post-mortem interval in the test group.

The regulation of β -catenin stability is a crucial control point for signalling through Wnt and perhaps other signal-transduction pathways⁹. To determine whether PS1 can affect the function of β -catenin by regulating its stability, we studied β -catenin degradation by using pulse–chase analysis. HEK293 cells were co-transfected with Myc-tagged β -catenin and either wild-type PS1 or PS1 containing mutations that cause familial Alzheimer's disease. β -Catenin was degraded with a half-life of about 1 hour in cells transfected with the vector alone, whereas it was stabilized over a 4-hour chase period in cells transfected with wild-type PS1 (Fig. 3a, b). In contrast, PS1 containing the M146V, C410Y or I143T mutations associated with familial Alzheimer's disease produced significantly less stabilization of β -catenin (Fig. 3a, b). The expression levels and rates of degradation of wild-type and mutant forms of PS1 were similar (data not shown). Thus, wild-type PS1 can stabilize β -catenin, whereas pathogenic PS1 mutants show loss of the stabilizing function.

We further studied the effect of PS1 mutations on β -catenin stability in PS1-transgenic mice. The electrophoretic migration of the human transgene-derived PS1 N-terminal fragment was different from that of the endogenous mouse fragment, allowing us to compare mice expressing similar levels of the transgene-derived PS1 fragment (Fig. 3c, lower panel). Analysis of β -catenin in the cerebral cortex of mice transgenic for PS1 containing the H163R or I143T mutations showed that a ladder of smaller β -catenin-degradation products, ranging from 36K to 75K, was markedly increased when compared with wild-type PS1 transgenics and non-transgenic littermate controls (Fig. 3c, upper panel). This pattern of increased

β -catenin-degradation products was detected with two different β -catenin antibodies and appeared in every transgenic animal studied from four different lines of mutant PS1 mice. The amount of full-length β -catenin did not show a consistent difference between control, wild-type PS1 and mutant PS1 transgenic mice (Fig. 3c). This may reflect the restriction of PS1 transgene expression to neurons, which prevents an effect on β -catenin in the more abundant glial cell population. These results indicate that PS1 mutations increase the degradation of β -catenin.

To investigate further the function of PS1 in the regulation of β -catenin stability, we studied β -catenin in fibroblast cultures derived from PS1-null mice¹². PS1-deficient fibroblasts had less full-length β -catenin than did wild-type controls (Fig. 3d, left panel). Furthermore, PS1-deficient fibroblasts showed a ladder of smaller β -catenin-degradation products that was not detected in control fibroblasts (Fig. 3d, right panel). These results indicate that the absence of PS1 results in increased degradation of β -catenin, and provide support for a physiological role of PS1 in the regulation of β -catenin stability.

We next determined whether β -catenin levels are altered in the brains of Alzheimer's disease patients with PS1 mutations. The mean level of β -catenin in the brains of patients with PS1 mutations was $46 \pm 20\%$ of that in controls (mean $\pm$ s.d., $P < 0.001$ by analysis of variance (ANOVA)) (Fig. 4a, b). In patients with sporadic Alzheimer's disease, the mean β -catenin concentration was slightly less than that of the control group, but this difference was not statistically significant (Fig. 4a, b). The same results were obtained when β -catenin levels were normalized to the endogenous levels of actin (data not shown). In contrast, the concentration of PS1, as indicated by immunoblot analysis of the PS1 N-terminal fragments, was not significantly different between patients with PS1 mutations and controls (Fig. 4a). Thus, pathogenic PS1 mutations are associated with significantly reduced β -catenin levels in the brains of patients with Alzheimer's disease.

We then determined whether reduced β -catenin levels increase neuronal degeneration by increasing neuronal vulnerability to amyloid- β protein¹³. Primary cultures of rat hippocampal neurons were co-transfected with β -catenin and green fluorescent protein (GFP) expression vectors, and transfected neurons were identified by double-label immunofluorescence for GFP and neuron-specific β -tubulin. We identified apoptotic neurons by nuclear staining with Hoechst dye. To inhibit β -catenin signalling, a dominant-negative β -catenin construct (Δ N/C- β -catenin), which inhibits β -catenin signalling through the Tcf/Lef family of transcription factors^{14,15}, was expressed. Expression of Δ N/C- β -catenin resulted in a marked increase in neuronal apoptosis induced by fibrillar amyloid- β protein; this increase in apoptosis was reversed by co-expression of wild-type β -catenin (Fig. 5a, b). Wild-type β -catenin and Δ N/C- β -catenin did not affect neuronal apoptosis in the absence of amyloid- β protein (Fig. 5b). Neuronal apoptosis induced by oxidative stress ($H_2O_2 + Fe$) was also increased by Δ N/C- β -catenin (data not shown). Thus, reduced β -catenin signalling increases neuronal vulnerability to apoptosis.

To confirm that the β -catenin signalling pathway is involved in the regulation of neuronal apoptosis, we studied the function of the transcription factor Tcf, which mediates β -catenin signalling¹⁴. To inhibit endogenous Tcf function, we used a dominant-negative Tcf construct (DN-Tcf-3) that inhibits Tcf signalling in several systems^{14,15}. Transfection of DN-Tcf-3 significantly increased neuronal apoptosis in both the absence and the presence of amyloid- β protein; this increase in apoptosis was reversed by co-expression of wild-type Tcf-3 (Fig. 5c). These results indicate that constitutive Tcf signalling protects against neuronal apoptosis, and that impaired signalling due to reduction in β -catenin levels can increase neuronal vulnerability to apoptosis.

These data show that β -catenin forms a complex with the PS1 N- and C-terminal fragments, leading to increased β -catenin stability.

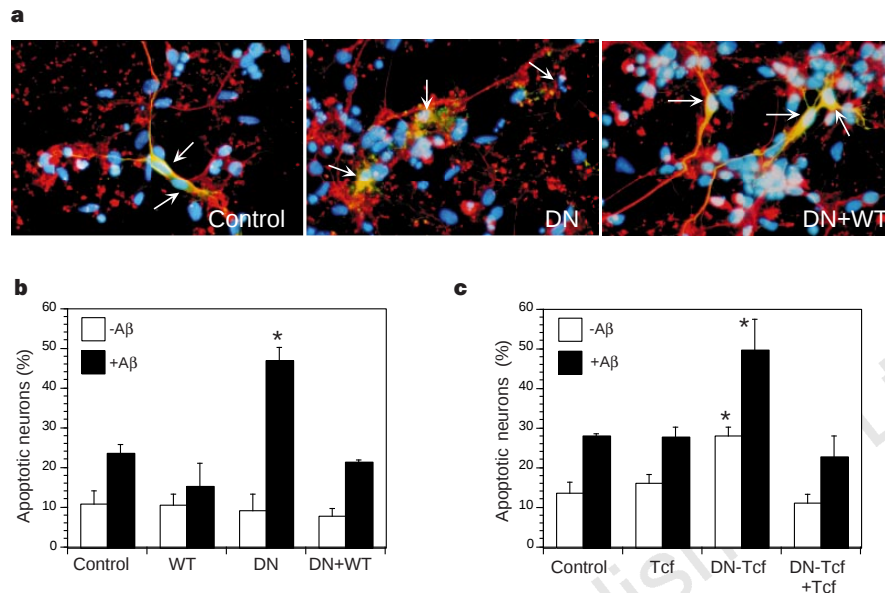


Figure 5 Loss of β -catenin signalling increases neuronal vulnerability to apoptosis. **a**, Hippocampal neurons were transfected with either the empty expression vector (control) or dominant-negative β -catenin (DN), or were co-transfected with dominant-negative and wild-type β -catenin (DN + WT). GFP was expressed to identify transfected cells. Following incubation with fibrillar amyloid- β protein (A β), the cultures were fixed and labelled with anti-GFP antibody (green), anti- β -tubulin antibody (red), and Hoechst dye (blue). Transfected neurons exhibit yellow cytoplasmic staining from the overlap of GFP and neuron-specific β -tubulin (arrows). Note the appearance of chromatin condensation and nuclear fragmentation in neurons expressing the dominant-negative β -catenin. **b**, Quan-

titative analysis of neuronal apoptosis induced by dominant-negative β -catenin. Apoptotic neurons were scored after incubation in the presence or absence of amyloid- β protein. **c**, Expression of dominant-negative Tcf-3 (DN-Tcf) increases neuronal apoptosis in both the absence and the presence of amyloid- β protein. Co-expression of wild-type Tcf-3 (Tcf) reverses the effect of DN-Tcf. Values represent the mean $\pm$ s.d.; $n = 4$ wells in a representative experiment. Asterisk indicates $P < 0.001$ relative to control by ANOVA with post-hoc Student-Neumann-Keuls test. The same statistical significance was obtained in five independent experiments.

The PS1 N- and C-terminal fragments can form a heteromeric complex^{6,16}. Pathogenic PS1 mutations increase β -catenin degradation *in vitro* and in transgenic mice, and are associated with reduced β -catenin levels in patients with Alzheimer's disease. PS1 mutations also alter the proteolytic processing of the amyloid precursor protein, giving rise to increased production of the 42-amino-acid form of amyloid- β protein^{12,17–20}. Thus, PS1 may function to regulate the proteolytic degradation of several different proteins.

Our results indicate that loss of signalling through the β -catenin–Tcf pathway increases neuronal vulnerability to apoptosis. These results are consistent with the observation that overexpression of APC, which negatively regulates β -catenin signalling, induces apoptosis in colon carcinoma cells²¹. A previous study obtained contrasting results, namely that Armadillo–Tcf signalling in *Drosophila* can induce apoptosis in developing retinal neurons²². This may reflect different effects of β -catenin–Tcf signalling in retinal versus hippocampal neurons, or in *Drosophila* versus mammalian cells. Increased apoptosis and neurotoxicity of amyloid- β protein have been observed in cells that express pathogenic PS1 mutations^{23–25}. PS1 mutations may increase neuronal vulnerability to apoptosis by altering the stability of β -catenin. Regulation of β -catenin signalling could therefore be a potential therapeutic approach to slowing the progression of neurodegeneration in Alzheimer's disease. □

Methods

Plasmids and antibodies. We cloned full-length PS1 into the expression vector pcDNA3, and the PS1 mutations M146V, C410, H163R and I143T were introduced by site-directed mutagenesis and confirmed by sequencing. The β -catenin, dominant-negative β -catenin (Δ N/C- β -catenin; amino acids 132–695), Tcf-3 and dominant-negative Tcf-3 (Δ N-Tcf-3) expression vectors have been described¹⁴. Rabbit polyclonal antisera against the PS1 N terminus and the PS1 C-terminal loop domain have been described⁷. Rabbit polyclonal and

mouse monoclonal antibodies against β -catenin (antibodies C2206 and C7207, respectively; Sigma; 1:250) were used for immunoprecipitation, and mouse monoclonal anti- β -catenin antibody (mAb14; Transduction Laboratories; 1:500) was used for immunoblotting. The anti-Myc-epitope antibody is clone 9E10.2 from ATCC; anti-SREBP1 antibody (C20) is from Santa Cruz Biotechnology; and anti-actin antibody (AC15) is from Sigma.

Immunoprecipitation, immunoblotting and immunofluorescence. We prepared cell lysates and human temporal cortex homogenates in NP-40 lysis buffer (1% NP-40, 10 mM HEPES, pH 7.5, 142.5 mM KCl, 5 mM MgCl₂, 1 mM EGTA) supplemented with protease inhibitors (Complete^R, Boehringer) by incubation for 1 h at 4°C with gentle rotation. Lysates were centrifuged at 14,000g for 30 min. The supernatant was then preincubated with Sepharose-4B beads for 2 h and adjusted for protein concentration. Immunoprecipitation was done by incubation with primary antibody and protein A- or protein G-Sepharose (Pharmacia) at 4°C for 12 h. The beads were washed by centrifugation and resuspension in 1% NP-40 buffer three times and then boiled for 5 min in reducing SDS sample buffer. The immunoprecipitated proteins were resolved by SDS-PAGE on a 4–20% polyacrylamide gel, and were electrotransferred to Immobilon-P membrane (Millipore). Immunoblots were developed with peroxidase-conjugated secondary antibody and enhanced chemiluminescence (ECL, Amersham). We analysed immunoprecipitates of proteins labelled with [³⁵S]methionine by phosphorimaging. Preabsorption antibody controls have been described⁷. For immunofluorescence microscopy, cells grown on coverslips were fixed in 3.7% paraformaldehyde, permeabilized in 0.4% Triton X-100, and blocked with 5% BSA in PBS.

Pulse-chase analysis. HEK293 cells in 6-cm dishes were transfected with complementary DNAs encoding Myc-tagged β -catenin (1 μ g) and wild-type or mutant presenilins, or the empty pcDNA3 expression vector (4 μ g). Forty-eight hours after transfection, cells were pulse-labelled with [³⁵S]methionine for 12 min and then chased in unlabelled medium containing a threefold excess of unlabelled methionine. Cells were lysed in 0.7 ml RIPA-DOC buffer (0.05 M Tris buffer, pH 7.2, 0.15 M NaCl, 1% Triton-X100, 1% deoxycholate, 0.1% SDS) supplemented with protease inhibitors and spun at 14,000g for 30 min at 4°C.

The supernatants were pre-cleared with Sepharose-4BL beads and immunoprecipitated with anti-Myc-epitope antibody (9E2.10). The immunoprecipitates were then washed twice with RIPA-DOC buffer, resolved by 4–20% SDS-PAGE, and quantified by phosphorimaging with subtraction of individual lane background.

Analysis of PS1-transgenic mice. Wild-type and mutant *PS1* cDNAs were cloned into a Thy-1 expression vector and transgenic mice were generated and bred as described²⁶. We identified founders by Southern blotting of genomic DNA from tail biopsies. Transgene-positive offspring were identified by polymerase chain reaction and confirmed by western blotting of cortical homogenates for human *PS1* fragments. Transgene expression was also monitored by *in situ* hybridization for the human *PS1* messenger RNA.

Quantitative analysis of β -catenin in the human brain. Grey matter was dissected from the temporal cerebral cortex of control cases ($n = 10$, average age 59 years, age range 25–86 years), cases with sporadic Alzheimer's disease ($n = 7$, average age 71, age range 63–82), and cases with familial Alzheimer's disease with *PS1* mutations (G209V; $n = 2$; C410Y, $n = 3$; M139I, $n = 2$; H163R, $n = 2$; average age 51, age range 41–59; one case with a C410Y mutation was 84 years old). The tissue was extracted in RIPA-DOC buffer supplemented with protease inhibitors and cleared by centrifugation (100,000g, 60 min, 4°C), and protein-equivalent samples were resolved by 4–20% SDS-PAGE. For detection of β -catenin, samples were heated for 1 min at 90°C before loading. For detection of *PS1*, samples were loaded without previous heating. The proteins were electrotransferred to Immobilon-P membrane and immunoblotted with anti-*PS1*-N (1:5,000) or anti- β -catenin (mAb14; 1:1,000), followed by incubation with ¹²⁵I-conjugated anti-rabbit or anti-mouse IgG and quantitative analysis by phosphorimaging.

Analysis of neuronal apoptosis. We established and transfected primary cultures of rat E18 hippocampal neurons as described²⁷. We added fibrillar amyloid- β protein residues 1–40 (ref. 28) to 40 μ M at 12 h after transfection, and incubated the cultures for a further 36 h. Transfected neurons were identified by double labelling for GFP and neuron-specific β -tubulin. Apoptotic neurons were identified by characteristic nuclear morphology after staining with Hoechst 33342 (1 μ g ml⁻¹).

Received 10 August; accepted 24 September 1998.

- Sherrington, R. *et al.* Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* **375**, 754–760 (1995).
- Levy-Lahad, E. *et al.* Candidate gene for chromosome 1 familial Alzheimer's disease locus. *Science* **269**, 973–977 (1995).
- Rogaev, E. I. *et al.* Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene. *Nature* **376**, 775–778 (1995).
- Li, J., Ma, J. & Potter H. Identification and expression of a potential familial Alzheimer disease gene on chromosome 1 related to AD3. *Proc. Natl Acad. Sci. USA* **92**, 12180–12184 (1995).
- Zhou, J. *et al.* Presenilin 1 interaction in the brain with a novel member of the Armadillo family. *Neuroreport* **8**, 1489–1494 (1997).
- Yu, G. *et al.* The presenilin 1 protein is a component of a high molecular weight complex that contains β -catenin. *J. Biol. Chem.* **273**, 164–16475 (1998).
- Busciglio, J. *et al.* Neuronal localization of presenilin-1 and association with amyloid plaques and neurofibrillary tangles in Alzheimer's disease. *J. Neurosci.* **17**, 5101–5107 (1997).
- Thinakaran, G. *et al.* Endoproteolysis of presenilin 1 and accumulation of processed derivatives *in vivo*. *Neuron* **17**, 181–190 (1996).
- Nusse, R. A versatile transcription effector of wingless signalling. *Cell* **89**, 321–323 (1997).
- Kovacs, D. M. *et al.* Alzheimer-associated presenilins 1 and 2: neuronal expression in brain and localization to intracellular membranes in mammalian cells. *Nature Med.* **2**, 224–229 (1996).
- Walter, J. *et al.* Alzheimer's disease-associated presenilins are differentially phosphorylated proteins located within the endoplasmic reticulum. *Mol. Med.* **2**, 673–691 (1996).
- De Strooper, B. *et al.* Deficiency of presenilin-1 inhibits the normal cleavage of amyloid precursor protein. *Nature* **391**, 387–390 (1998).
- Yankner, B. A. Mechanisms of neuronal degeneration in Alzheimer's disease. *Neuron* **16**, 921–932 (1996).
- Molenaar, M. *et al.* Xctf-3 transcription factor mediates β -catenin-induced axis formation in *Xenopus* embryos. *Cell* **86**, 391–399 (1996).
- van de Wetering, M. *et al.* Armadillo coactivates transcription driven by the product of the *Drosophila* segment polarity gene *TCF*. *Cell* **88**, 789–799 (1997).
- Capell, A. *et al.* The proteolytic fragments of the Alzheimer's disease-associated presenilin-1 form heterodimers and occur as a 100–150-kDa molecular mass complex. *J. Biol. Chem.* **273**, 3205–3211 (1998).
- Scheuner, D. *et al.* Secreted amyloid beta-protein similar to that in the senile plaques of Alzheimer's disease is increased *in vivo* by the presenilin 1 and 2 and APP mutations linked to familial Alzheimer's disease. *Nature Med.* **2**, 864–870 (1996).
- Duff, K. *et al.* Increased amyloid- β 42(43) in brains of mice expressing presenilin 1. *Nature* **383**, 710–713 (1996).
- Borchelt, D. R. *et al.* Familial Alzheimer's disease-linked presenilin-1 variants elevate A β 1–42/1–40 ratio *in vitro* and *in vivo*. *Neuron* **17**, 1005–1013 (1996).
- Citron, M. *et al.* Mutant presenilins of Alzheimer's disease increase production of 42-residue amyloid beta-protein in both transfected cells and transgenic mice. *Nature Med.* **3**, 67–72 (1997).
- Morin, P. J., Vogelstein, B. & Kinzler, K. W. Apoptosis and APC in colorectal tumorigenesis. *Proc. Natl Acad. Sci. USA* **93**, 7950–7954 (1996).

- Ahmed, Y., Hayashi, S., Levine, A. & Wiesthaus, E. Regulation of Armadillo by a *Drosophila* APC inhibits neuronal apoptosis during retinal development. *Cell* **93**, 1171–1182 (1998).
- Wolozin, B. *et al.* Participation of presenilin 2 in apoptosis: enhanced basal activity conferred by an Alzheimer mutation. *Science* **274**, 1710–1713 (1996).
- Guo, Q. *et al.* Alzheimer's presenilin mutation sensitizes neural cells to apoptosis induced by trophic factor withdrawal and amyloid β -peptide: involvement of calcium and oxyradicals. *J. Neurosci.* **17**, 4212–4222 (1997).
- Guo, Q. *et al.* Par-4 is a mediator of neuronal degeneration associated with the pathogenesis of Alzheimer disease. *Nature Med.* **4**, 957–962 (1998).
- Sturchler-Pierrat, C. *et al.* Two amyloid precursor protein transgenic mouse models with Alzheimer disease-like pathology. *Proc. Natl Acad. Sci. USA* **94**, 13287–13292 (1997).
- Dudek, H. *et al.* Regulation of neuronal survival by the serine-threonine protein kinase Akt. *Science* **275**, 661–665 (1997).
- Lorenzo, A. & Yankner, B. A. Beta-amyloid neurotoxicity requires fibril formation and is inhibited by Congo Res. *Proc. Natl Acad. Sci. USA* **91**, 12243–12247 (1994).

Acknowledgements. We thank D. Nochlin, C. Lipka, T. Bird, C. Rosenberg, A. Roses, D. Pollin and J. Rogers for autopsy human brain tissue; K. Burki and B. Lederman for assistance in the generation of transgenic mice; and Y. Sun for discussions. This work was supported by grants from the NIH, the Alzheimer's Association and Novartis Pharma Ltd (to B.A.Y.), an NIH training grant and a fellowship from The Medical Foundation (to Z.Z.), a fellowship from the Deutsche Forschungsgemeinschaft (to H.H.), a Pew Scholarship (to X.H.), and an NIH MRRC Core Grant.

Correspondence and requests for materials should be addressed to B.A.Y. (e-mail: Yankner@A1.tch.harvard.edu).

Cripto is required for correct orientation of the anterior–posterior axis in the mouse embryo

Jixiang Ding*, Lu Yang*, Yu-Ting Yan*, Amy Chen†, Nishita Desai*, Anthony Wynshaw-Boris† & Michael M. Shen*

* Center for Advanced Biotechnology and Medicine and Dept of Pediatrics, UMDNJ-Robert Wood Johnson Medical School, 679 Hoes Lane, Piscataway, New Jersey 08854, USA

† Laboratory of Genetic Disease Research, National Human Genome Research Institute, 49 Convent Drive, Bethesda, Maryland 20892, USA

The anterior–posterior axis of the mouse embryo is established by two distinct organizing centres in the anterior visceral endoderm and the distal primitive streak^{1–7}. These organizers induce and pattern the head and trunk respectively, and have been proposed to be localized through coordinate cell movements that rotate a pre-existing proximal–distal axis^{6,8}. Here we show that correct localization of both head- and trunk-organizing centres requires *Cripto*^{9,10}, a putative signalling molecule that is a member of the EGF-CFC gene family^{11,12}. Before gastrulation, *Cripto* is asymmetrically expressed in a proximal–distal gradient in the epiblast, and subsequently is expressed in the primitive streak and newly formed embryonic mesoderm. A *Cripto* null mutation generated by targeted gene disruption results in homozygous *Cripto*^{-/-} embryos that mostly consist of anterior neuroectoderm and lack posterior structures, thus resembling a head without a trunk. Notably, markers of the head organizer are located at the distal end of the embryo, whereas markers of the primitive streak are absent or localized to the proximal side. Our results indicate that *Cripto* signalling is essential for the conversion of a proximal–distal asymmetry into an orthogonal anterior–posterior axis.

Members of the EGF-CFC gene family, including *Cripto*, murine *Cryptic*¹¹, *Xenopus FRL-1*¹³, and zebrafish *one-eyed pinhead*¹², encode secreted proteins that contain a divergent epidermal growth factor (EGF)-like motif and a novel cysteine-rich CFC motif^{11,12}. EGF-CFC genes are expressed during gastrulation, with limited expression at later embryonic stages, and have been implicated in early development. For example, *FRL-1* has mesoderm and neural-inducing activities in *Xenopus* animal cap assays¹³, and *one-eyed pinhead* is required for prechordal plate formation during zebrafish gastrulation¹². Expression of *Cripto* can be detected in the epiblast

of pre-streak- and primitive-streak-stage embryos^{9,10}; *Cripto* is necessary for cardiogenesis during embryonic stem cell differentiation in culture¹⁴. We demonstrate here that *Cripto* is required for early embryogenesis in the mouse, thus establishing an essential role for the EGF-CFC family in mammalian development.

We have found that the expression of *Cripto* during pregastrulation and gastrulation stages is dynamic and is associated with early signs of overt anterior–posterior asymmetry. Before gastrulation, at 5.5 days post coitum (d.p.c.), *Cripto* expression is initially symmetric and uniform in the epiblast, and is not found in the extra-embryonic visceral endoderm (Fig. 1a, b). At 6.0 d.p.c., before primitive streak formation, *Cripto* becomes asymmetrically expressed in the epiblast, in a graded proximal–distal distribution (Fig. 1c–e). By the onset of gastrulation (6.5 d.p.c.), *Cripto* expression localizes to the region of the nascent primitive streak (Fig. 1f–h). In the early stages of gastrulation (6.75 d.p.c.), *Cripto* expression is found in the primitive streak and in the wings of embryonic mesoderm that spread rostrally, while regressing caudally in the epiblast (Fig. 1i, j). During the early neural plate stage (7.25 d.p.c.), expression of *Cripto* fades in the embryonic mesoderm and becomes increasingly confined to the primitive streak and head process (Fig. 1k, l). *Cripto* expression disappears completely by the late neural plate stage (7.75 d.p.c.), whereas a later phase of expression in cardiac progenitors initiates at the late head-fold stage (8.0 d.p.c.) (J.D., H. Wang and M.M.S., unpublished results).

To investigate whether *Cripto* activity is required for proper gastrulation, we generated a null mutation by targeted gene dis-

ruption, simultaneously ‘knocking-in’ a promoterless *lacZ* marker gene (Fig. 2a). The resulting heterozygous *Cripto*^{+/-} mice are phenotypically normal, and *Cripto*^{+/-} embryos display β -galactosidase staining patterns that resemble those obtained by *in situ* hybridization, albeit with a slight lag in developmental stage that probably reflects protein persistence (Fig. 1m–r). However, homozygosity for the *Cripto* targeted mutation results in embryonic lethality (Fig. 2b). Although no homozygous embryos were obtained after 10.5 d.p.c., *Cripto* homozygotes were recovered in the expected mendelian ratios at 6.75 d.p.c. through to 8.5 d.p.c. (Fig. 2c).

Morphological and histological analysis demonstrated that *Cripto* homozygotes have striking defects in embryonic mesoderm formation and axial organization. The earliest stage at which *Cripto* homozygotes can be reliably distinguished is 6.75 d.p.c., when wild-type littermates are at early-to-mid-streak stages of gastrulation. At this stage, *Cripto* mutants can be identified by the absence of a primitive streak and embryonic mesoderm, and by a thickened layer of visceral endoderm at the distal tip of the egg cylinder (Fig. 2d,h,i). In contrast, in wild-type embryos, the visceral endoderm is thin distally, but is thicker proximally towards the embryonic/extra-embryonic constriction, particularly on the anterior side² (Fig. 2d, g). Later, at 7.5 and 8.5 d.p.c. *Cripto* mutants appear to be completely deficient for embryonic mesoderm derivatives such as somites and cardiac tissue, whereas extra-embryonic mesoderm is often formed, including the allantois and blood islands of the visceral yolk sac mesoderm (Fig. 2d–l). Notably, *Cripto* mutant

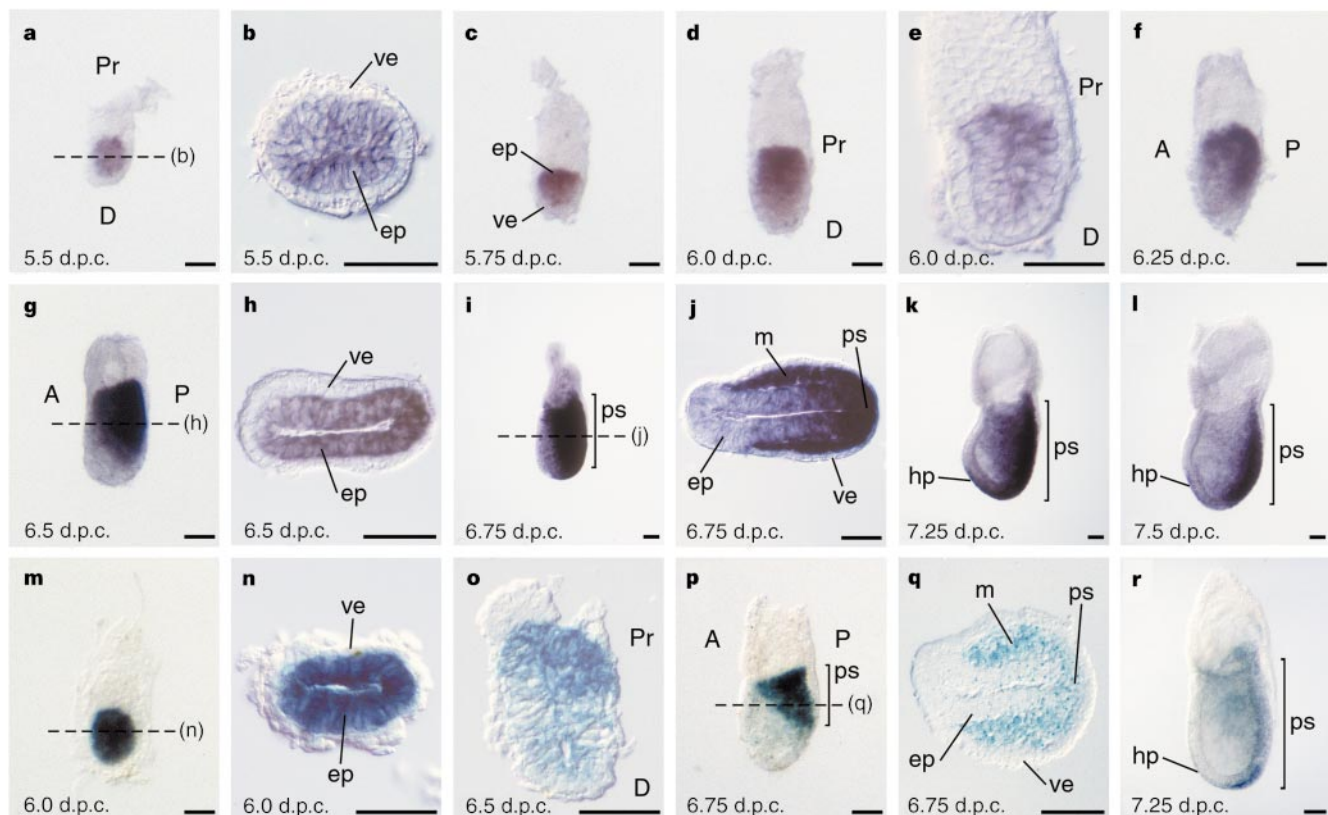


Figure 1 Expression of *Cripto* at pre-gastrulation and gastrulation stages. **a–l**, Whole-mount *in situ* hybridization analysis. **a**, Uniform symmetric staining in the epiblast at 5.5 d.p.c.; **b**, a cross section shows lack of expression in the visceral endoderm. **c, d**, Proximal–distal gradient of expression in the epiblast. **e**, Sagittal section of **d**. **f, g**, Expression shifts caudally before the onset of gastrulation at 6.5 d.p.c. **h**, Cross-section of **g** shows widespread expression in the epiblast and no expression in the visceral endoderm. **i**, Mid-streak stage; **j**, cross-section shows intense staining in the newly formed embryonic mesoderm. **k, l**, Expression persists in the primitive streak and head process at the neural plate

stage. **m–r**, β -galactosidase staining of *Cripto* heterozygotes. **m, n**, Uniform staining in the epiblast before gastrulation. **o**, Sagittal section shows proximal–distal graded staining just before gastrulation. **p**, Early-streak stage embryo, and **q**, cross-section. **r**, Early neural-plate-stage embryo: note more intense staining at distal end of the primitive streak. In all panels, anterior faces to the left when anterior–posterior orientation can be identified; staging before primitive streak formation is approximate. Scale bars, 0.05 mm. A, anterior; D, distal; ep, epiblast; hp, head process; m, mesoderm; P, posterior; ps, primitive streak (bar denotes extent of streak); ve, visceral extra-embryonic endoderm.

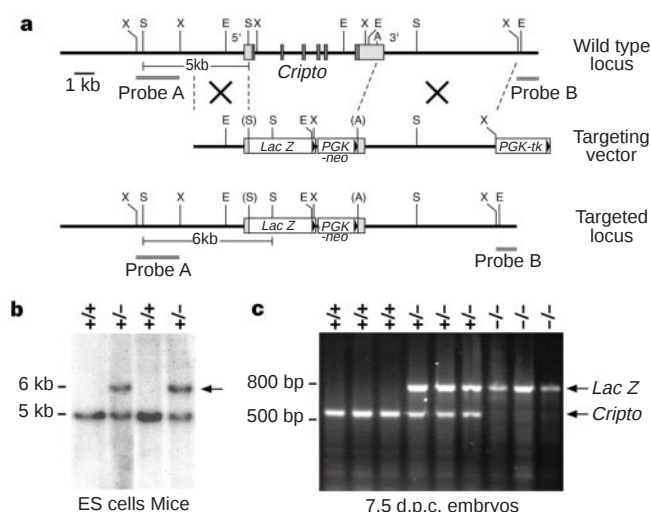
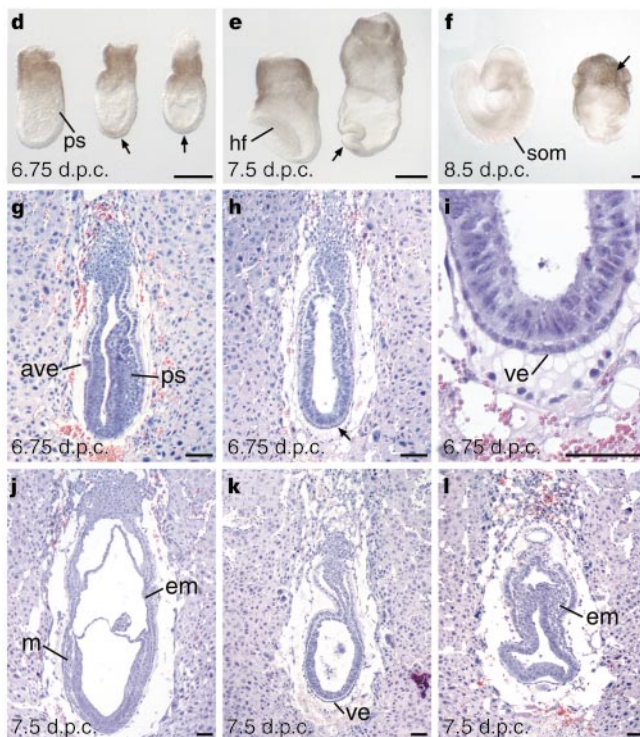


Figure 2 Targeted disruption of *Cripto* and characterization of the null phenotype. **a**, Homologous recombination with the targeting vector deletes the entire *Cripto* coding region (dark boxes), and inserts a promoter-less *LacZ* marker gene into the 5' untranslated region; triangles indicate orientation of *PGK-neo*, *LacZ*, and *PGK-tk* cassettes. A, *ApaI*; E, *EcoRI*; S, *SspI*; X, *XbaI*. **b**, Southern blotting using a 5' flanking probe (A) detects a 5-kb *SspI* fragment in wild-type genomic DNA and a 6-kb fragment (arrow) from the targeted allele in ES cells (lanes 1, 2) and in mice (lanes 3, 4). Targeted ES clones were confirmed using a *LacZ* probe and a 3' external probe (B) (not shown). **c**, PCR analysis of extra-embryonic tissues demonstrates recovery of wild-type (lanes 1–3), heterozygous (lanes 4–6), and homozygous mutant embryos (lanes 7–9) at 7.5 d.p.c.; the primers amplify a 751-bp *LacZ* fragment and a 536-bp *Cripto* genomic fragment deleted in the targeted allele. **d–f**, Morphology of dissected embryos; scale bars represent 0.2 mm. **d**, Wild-type mid-streak stage embryo at 6.75 d.p.c. (left) and two *Cripto* mutant littermates (centre, right), showing thickened visceral endoderm at the distal tip

embryos have ectopic folds of ectodermal tissue that are located distally, and frequently lack an overt anterior–posterior axis (Fig. 2d–l).

We further characterized the *Cripto* mutant phenotype by examining expression of appropriate marker genes for specific tissues in embryos dissected at 7.5 d.p.c. and 8.5 d.p.c. Consistent with our histological findings, we observed no embryonic expression of *HNF-3 β* and *Sonic hedgehog* (*Shh*), which are markers of axial mesoderm and definitive endoderm at these stages (Fig. 3a, b). We also found no evidence for formation of paraxial mesoderm using a *Mox1* probe¹⁵ (Fig. 3c). Similar results were obtained using a *Lim1* probe for caudal lateral plate mesoderm (data not shown). To determine the identity of the ectopic ectodermal folds in *Cripto* mutants, we next investigated the expression of several neuroectoderm markers. In all *Cripto* homozygotes examined, we found widespread expression in the ectoderm layer for *Otx2*, which is initially expressed throughout the undifferentiated epiblast before gastrulation and then marks the prospective forebrain and midbrain at 7.5 d.p.c.¹⁶ (Fig. 3d, e). In addition, expression of the forebrain marker *BF1*¹⁷ was observed in four out of seven embryos at 7.5 d.p.c. and 8.5 d.p.c. (Fig. 3f). *En1*, a marker of prospective midbrain and anterior hindbrain, was expressed in every homozygous embryo examined, usually in an asymmetric patch (Fig. 3g, h). In contrast to



these anterior neuroectoderm markers, staining was not observed for *Krox20*, a marker for rhombomeres 3 and 5 of the posterior hindbrain (Fig. 3i). Similarly, expression of the more caudal neural and mesodermal markers *HoxB1* and *HoxB4* was completely absent (Fig. 3j; data not shown). To rule out the possibility that the tissues in *Cripto* mutant embryos correspond to undifferentiated epiblast, we examined expression of the epiblast marker *Oct4* and found no evidence of expression at 7.5 d.p.c. (data not shown). Based on these data, we conclude that *Cripto* mutant embryos are severely deficient for production of embryonic mesoderm and endoderm, and instead largely consist of anterior neuroectoderm.

The absence of embryonic mesoderm and ectopic neuroectoderm in *Cripto* mutants suggested that formation and/or localization of the trunk and head organizing centres might be defective. To determine whether the trunk organizing centre is affected, we used *Brachyury*, an early marker of the primitive streak, and found that it was expressed in a proximal band near the embryonic/extraembryonic constriction at 6.75 d.p.c. (Fig. 3k, l), but was undetectable at 7.5 d.p.c. (data not shown). To confirm this finding, we examined the markers *Fgf8* and *Evx1*, which mark the entire streak and caudal (proximal) primitive streak respectively¹⁸, and obtained similar results at 6.75 d.p.c. (Fig. 3m; data not shown). We also examined expression of *Lim1*, which marks both the primitive

Finally, we investigated whether head organizing activities are affected in *Cripto* mutants by examining several markers that are normally expressed in the anterior visceral endoderm during early gastrulation. Expression of *Hesx1* (also known as *Rpx*), which marks the anterior visceral endoderm and is subsequently induced in the adjacent neuroectoderm^{1,19}, was found throughout the ectodermal layer of the distal region of the egg cylinder at 6.75 d.p.c. (Fig. 3p, q); this ectopic expression significantly precedes the normal time at

In summary, our analysis of *Cripto* mutants indicates that early markers of a head-organizing centre (*Hex*, *Cerberus-like*) are mislocalized in the distal visceral endoderm and, as a result, anterior neuroectoderm is induced distally. In contrast, early markers of the primitive streak (*Brachyury*, *Fgf8*) are mislocalized proximally, whereas little or no expression is observed for later markers of the distal primitive streak (*Goosecoid*, *HNF-3 β*), which corresponds to the position of the trunk organizing centre and future node. These

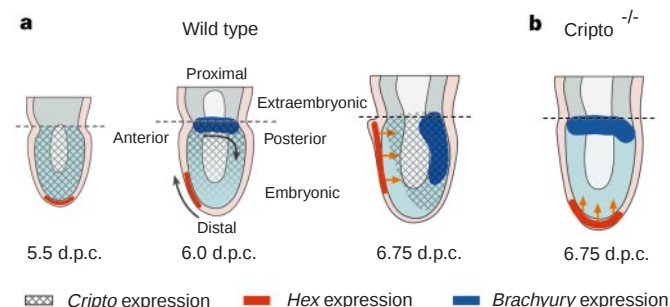


Figure 4 Schematic interpretation of the *Cripto* mutant phenotype. **a**, Model for anterior-posterior axis formation proposed by Beddington *et al.* (adapted from refs 6, 8). Expression of *Hex* (red) and *Brachyury* (blue) correspond to the position of prospective anterior and posterior organizing centres, respectively; orange arrows represent putative head-inducing activities. Also shown is the expression of *Cripto* (hatched) at these stages, based on Fig. 1. Dashed lines indicate the embryonic/extra-embryonic boundary. **b**, Deduced locations of anterior and posterior organizing centres in *Cripto* mutant embryos. We propose that mislocalization of head-organizing activities accounts for the ectopic generation of anterior neuroectoderm in a distal position. Note that loss of *Cripto* activity results in a phenotype in the visceral endoderm, where it is not expressed, consistent with its putative role as a signalling factor, although the basis for this effect may be indirect.

findings correlate with the failure to form embryonic mesoderm, but not extra-embryonic mesoderm, and demonstrate that anterior neural fates can arise in the absence of embryonic mesoderm. Moreover, the lack of posterior neuroectoderm in *Cripto* mutants is consistent with a lack of signals from paraxial mesoderm that are normally required to caudalize neuroectoderm²⁰. Consequently, *Cripto* mutant embryos display a 'head-without-trunk' phenotype that is superficially the opposite of that found in mutants for *Otx2* and *Lim1*, which have defects in head-organizing activities^{21–23}. Mislocalized distal expression of *Cerberus*-like has been observed in *Otx2* mutants⁵, and mislocalized proximal expression of *Gooseoid* in *Otx2*²² and *Lim-1* mutants²¹; however, these mutants have at least partially intact trunk-organizing activities. Furthermore, the *Cripto* null phenotype also differs from that of *Smad2*, which results in a lack of anterior-posterior asymmetry and no localized expression of markers for both head and trunk organizing centres⁷. Thus, *Smad2* is required for the generation of the anterior-posterior organizing centres, whereas *Cripto* is required for their correct localization.

Our results support a recent model proposed to explain the formation of the anterior-posterior axis in the mouse embryo. Beddington *et al.* have suggested that precursors of the head- and trunk-organizing centres originate on the distal and proximal ends of the pregastrulation embryo, and that vectorial cell movements convert this pre-existing distal-proximal asymmetry into an anterior-posterior axis before gastrulation^{6,8} (Fig. 4a). Our findings agree with this model as we found that *Cripto* is required for the correct orientation of the anterior-posterior axis. In particular, our data indicate that rotation of the anterior-posterior axis fails to occur in the absence of *Cripto* activity, resulting in mislocalization of the head- and trunk-organizing centres (Fig. 4b). We suggest that *Cripto* signalling, possibly involving its proximal-distal gradient of expression, is required for the vectorial movements that correctly orient the anterior-posterior axis, and perhaps for coordinating these cell movements in two distinct tissue layers.

Relatively little is known about the signalling activities of EGF-CFC protein products and the molecular identities of their putative receptor(s), although recombinant *Cripto* protein activates the Ras/Raf/MAPK pathway in cultured mammary epithelial cells²⁴. Our demonstration of an essential role for *Cripto* in anterior-posterior

axis formation underscores the importance of this newly identified family of signalling molecules in vertebrate development. □

Methods

Gene targeting. A partial cDNA corresponding to base pairs (bp) 41 to 985 of the published *Cripto* sequence⁹ was amplified by RT-PCR from F9 teratocarcinoma cells, and used to screen a λFIXII library constructed from 129Sv/J genomic DNA (Stratagene). Positive phage clones were screened by PCR to distinguish the authentic *Cripto* locus from pseudogenes²⁵. To construct a targeting vector for *Cripto*, a 6.5-kb *ApaI* to *EcoRI* fragment was cloned into the *BamHI/EcoRI* sites of *pPNT*²⁶ to create the 3' flank. The 5' flank was cloned as a 2.6-kb *XhoI* to *SspI* fragment cloned into the *NotI/XhoI* sites of *pPNT* containing the 3' flank. Note that the *SspI* site of the 5' arm is located at bp 175 of the *Cripto* cDNA sequence in the 5' untranslated region. The linearized construct was electroporated into TC1 ES cells²⁸; targeted clones were obtained at a frequency of 13% (19/145). ES cell culture and blastocyst injection were done as described²⁹. Chimaeric males obtained following blastocyst injection were bred with Black Swiss females (Taconic), and germline transmission was obtained from one targeted ES clone. The targeted mutation has been maintained through backcrossing with outbred Black Swiss mice, and on an inbred 129/SvEv strain background; the phenotype of *Cripto* mutant embryos appears similar in both genetic backgrounds.

Embryo genotyping and phenotypic analysis. Staging of embryos was done according to standard criteria³⁰; 0.5 d.p.c. is considered to be noon of the day of the copulatory plug. Embryos analysed at 7.5 d.p.c. and 8.5 d.p.c. were genotyped by PCR using extraembryonic tissues; for 6.75 d.p.c., over 50 embryos have been genotyped, with a strict correlation observed between the genotype and the phenotype described. PCR amplification for genotyping was done at an annealing temperature of 59°C, using the following primers: for *LacZ*, 5' CCGCGCTGTACTGGAGGCTGAAG 3' and 5' ATACTGCACCGGGCGGGAAGGAT 3'; for *Cripto*, 5' GCGCAGCTTCCAACCTCAATC 3' and 5' TTCCAAGGCAACCAGGGCTACAC 3' (corresponding to bp 2,464–2,999 of the genomic sequence²⁵). Histological analysis of intact decidua was performed as described²⁹.

In situ hybridization and probes. Whole-mount *in situ* hybridization was carried out as previously described¹¹ using the indicated probes, except that embryos were processed in 0.74-μm mesh inserts (Costar) for ease of handling⁶. For each probe, at least five, and usually six or more, *Cripto* mutant embryos were examined at each developmental stage analysed. Whole-mount embryos were stained for β-galactosidase activity as described²⁷.

Received 8 June; accepted 31 July 1998.

1. Thomas, P. & Beddington, R. S. P. Anterior primitive endoderm may be responsible for patterning the anterior neural plate in the mouse embryo. *Curr. Biol.* **6**, 1487–1496 (1996).
2. Varlet, L., Collignon, J. & Robertson, E. J. *nodal* expression in the primitive endoderm is required for specification of the anterior axis during mouse gastrulation. *Development* **124**, 1033–1044 (1997).
3. Belo, J. A. *et al.* *Cerberus*-like is a secreted factor with neutralizing activity expressed in the anterior primitive endoderm of the mouse gastrula. *Mech. Dev.* **68**, 45–57 (1997).
4. Tam, P. P. & Behringer, R. R. Mouse gastrulation: the formation of a mammalian body plan. *Mech. Dev.* **68**, 3–25 (1997).
5. Biben, C. *et al.* Murine cerberus homologue mCerber-1: a candidate anterior patterning molecule. *Dev. Biol.* **194**, 135–151 (1998).
6. Thomas, P. Q., Brown, A. & Beddington, R. S. P. *Hex*: homeobox gene revealing peri-implantation asymmetry in the mouse embryo and an early transient marker of endothelial cell precursors. *Development* **125**, 85–94 (1998).
7. Waldrup, W. R., Bikoff, E. K., Hoodless, P. A., Wrana, J. L. & Robertson, E. J. *Smad2* signaling in extraembryonic tissues determines anterior-posterior polarity of the early mouse embryo. *Cell* **92**, 797–808 (1998).
8. Beddington, R. S. P. & Robertson, E. J. Anterior patterning in mouse. *Trends Genet.* **14**, 277–284 (1998).
9. Dono, R. *et al.* The murine *cripto* gene: expression during mesoderm induction and early heart morphogenesis. *Development* **118**, 1157–1168 (1993).
10. Johnson, S. E., Rothstein, J. L. & Knowles, B. B. Expression of epidermal growth factor family gene members in early mouse development. *Dev. Dyn.* **201**, 216–226 (1994).
11. Shen, M. M., Wang, H. & Leder, P. A differential display strategy identifies *Cryptic*, a novel EGF-related gene expressed in the axial and lateral mesoderm during mouse gastrulation. *Development* **124**, 429–442 (1997).
12. Zhang, J., Talbot, W. S. & Schier, A. F. Positional cloning identifies zebrafish *one-eyed pinhead* as a permissive EGF-related ligand required during gastrulation. *Cell* **92**, 241–251 (1998).
13. Kinoshita, N., Minshall, J. & Kirschner, M. W. The identification of two novel ligands of the FGF receptor by a yeast screening method and their activity in *Xenopus* development. *Cell* **83**, 621–630 (1995).
14. Xu, C., Liguori, G., Adamson, E. D. & Persico, M. G. Specific arrest of cardiogenesis in cultured embryonic stem cells lacking *Cripto-1*. *Dev. Biol.* **196**, 237–247 (1998).
15. Candia, A. F. *et al.* *Mox-1* and *Mox-2* define a novel homeobox gene subfamily and are differentially expressed during early mesodermal patterning in mouse embryos. *Development* **116**, 1123–1136 (1992).

16. Ang, S. L., Conlon, R. A., Jin, O. & Rossant, J. Positive and negative signals from mesoderm regulate the expression of mouse *Otx2* in ectoderm explants. *Development* **120**, 2979–2989 (1994).
17. Tao, W. & Lai, E. Telencephalon-restricted expression of *BF-1*, a new member of the HNF-3/fork head gene family, in the developing rat brain. *Neuron* **8**, 957–966 (1992).
18. Crossley, P. H. & Martin, G. R. The mouse *Fgf8* gene encodes a family of polypeptides and is expressed in regions that direct outgrowth and patterning in the developing embryo. *Development* **121**, 439–451 (1995).
19. Hermesz, E., Macken, S. & Mahon, K. A. *Rpx*: a novel anterior-restricted homeobox gene progressively activated in the prechordal plate, anterior neural plate and Rathke's pouch of the mouse embryo. *Development* **122**, 41–52 (1996).
20. Muhr, J., Jessell, T. M. & Edlund, T. Assignment of early caudal identity to neural plate cells by a signal from caudal paraxial mesoderm. *Neuron* **19**, 487–502 (1997).
21. Shawlot, W. & Behringer, R. R. Requirement for *Lim1* in head-organizer function. *Nature* **374**, 425–430 (1995).
22. Ang, S. L. *et al.* A targeted mouse *Otx2* mutation leads to severe defects in gastrulation and formation of axial mesoderm and to deletion of rostral brain. *Development* **122**, 243–252 (1996).
23. Rhinn, M. *et al.* Sequential roles for *Otx2* in visceral endoderm and neuroectoderm for forebrain and midbrain induction and specification. *Development* **125**, 845–856 (1998).
24. Kannan, S. *et al.* *Cripto* enhances the tyrosine phosphorylation of *Shc* and activates mitogen-activated protein kinase (MAPK) in mammary epithelial cells. *J. Biol. Chem.* **272**, 3330–3335 (1997).
25. Liguori, G. *et al.* Characterization of the mouse *Tdgl* gene and *Tdgl* pseudogenes. *Mamm. Genome* **7**, 344–348 (1996).
26. Tybulewicz, V. L., Crawford, C. E., Jackson, P. K., Bronson, R. T. & Mulligan, R. C. Neonatal lethality and lymphopenia in mice with a homozygous disruption of the *c-abl* proto-oncogene. *Cell* **65**, 1153–1163 (1991).
27. Shalaby, F. *et al.* Failure of blood-island formation and vasculogenesis in *Flk-1* deficient mice. *Nature* **376**, 62–66 (1995).
28. Deng, C., Wynshaw-Boris, A., Zhou, F., Kuo, A. & Leder, P. Fibroblast growth factor receptor 3 is a negative regulator of bone growth. *Cell* **84**, 911–921 (1996).
29. Deng, C.-X. *et al.* Murine *FGFR-1* is required for early postimplantation growth and axial organization. *Genes Dev.* **8**, 3045–3057 (1994).
30. Downs, K. M. & Davies, T. Staging of gastrulating mouse embryos by morphological landmarks in the dissecting microscope. *Development* **118**, 1255–1266 (1993).

Acknowledgements. We thank C. Abate-Shen, R. Beddington, R. Behringer, E. DeRobertis, A. Joyner, R. Krumlauf, E. Lai, G. Martin, A. McMahon, J. Rossant and C. Wright for gifts of probes and reagents; H. Wang and L. Garrett for technical assistance; and C. Abate-Shen, R. Steward, and members of M.M.S.'s laboratory for advice and comments on the manuscript. This work was supported by a postdoctoral fellowship from the American Heart Association (J.D.), and by grants from the American Heart Association, the U.S. Army Breast Cancer Research Program, and the NIH (to M.M.S.).

Correspondence and requests for materials should be addressed to M.M.S. (e-mail: mshen@cabm.rutgers.edu).

Sexually dimorphic development of the mammalian reproductive tract requires *Wnt-7a*

Brian A. Parr* & Andrew P. McMahon

Department of Molecular and Cellular Biology, The Biolabs, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA

An important feature of mammalian development is the generation of sexually dimorphic reproductive tracts from the Müllerian and Wolffian ducts. In females, Müllerian ducts develop into the oviduct, uterus, cervix and upper vagina, whereas Wolffian ducts regress. In males, testosterone promotes differentiation of Wolffian ducts into the epididymis, vas deferens and seminal vesicle. The Sertoli cells of the testes produce Müllerian-inhibiting substance, which stimulates Müllerian duct regression in males^{1–4}. The receptor for Müllerian-inhibiting substance is expressed by mesenchymal cells underlying the Müllerian duct that are thought to mediate regression of the duct^{5–7}. Mutations that inactivate either Müllerian-inhibiting substance or its receptor allow development of the female reproductive tract in males^{8–12}. These pseudohermaphrodites are frequently infertile because sperm passage is blocked by the presence of the female reproductive system^{9,10,12}. Here we show that male mice lacking the signalling molecule *Wnt-7a* fail to undergo regression of the Müllerian duct as a result of the absence of the receptor for Müllerian-inhibiting substance. *Wnt-7a*-deficient females are infertile because of abnormal development of the oviduct and uterus, both of which are

Müllerian duct derivatives. Therefore, we propose that signalling by *Wnt-7a* allows sexually dimorphic development of the Müllerian ducts.

Wnt-7a is a member of the *Wnt* family of secreted glycoproteins, which function as signalling molecules in several developmental contexts¹³. We have described previously the generation of a mutant *Wnt-7a* allele by homologous recombination in mouse embryonic stem cells¹⁴. Mice homozygous for the targeted allele exhibit defects in limb patterning, but are otherwise fully viable^{14,15}. However, we did not observe pregnancies when homozygous *Wnt-7a*-mutant males or females were mated.

To study the *Wnt-7a*-associated sterility, we dissected the reproductive tracts of postnatal mice. The Müllerian duct derivatives of homozygous *Wnt-7a*-mutant neonate or adult females, as compared with their wild-type or heterozygous littermates, failed to differentiate properly. At the neonatal stage, Müllerian ducts are present but there are no signs of coiled oviducts in homozygous *Wnt-7a*-mutant females (Fig. 1a–c). In addition, the uterine wall is thinner and notably less muscular than that of the wild-type mice (Fig. 1c). Adult *Wnt-7a*^{−/−} females still lack visibly coiled oviducts,

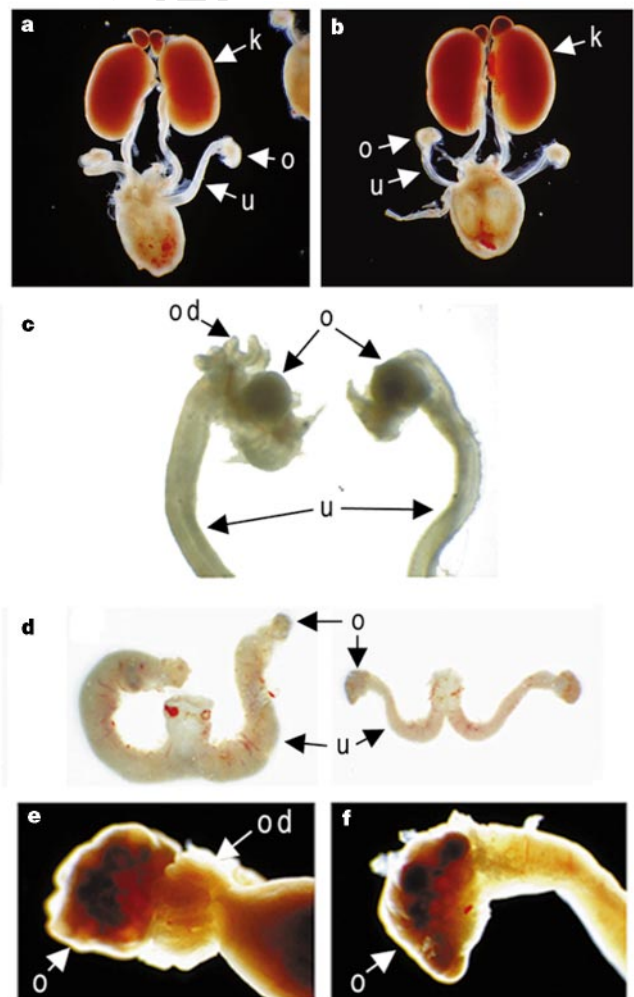


Figure 1 Development of female reproductive tracts in wild-type and *Wnt-7a*-mutant mice. **a, b**, The urogenital region of **a**, wild-type and **b**, mutant neonates. k, kidney; o, ovary; u, uterus. **c**, Coiled oviducts (od) are seen in wild-type (left) but not in mutant (right) neonatal mice. The uterus of the mutants is less muscular and more slender than the wild-type uterus. **d**, Adult reproductive tracts of wild-type (left) and homozygous *Wnt-7a*-mutant (right) females. The wild-type uterus is substantially larger and thicker than the mutant uterus. However, ovarian development is similar. **e, f**, The coiled oviduct, which is visible in wild-type mice (**e**), is not apparent in *Wnt-7a* mutants (**f**). Comparative photographs of wild-type and mutant siblings in all figures are taken at the same magnification.

* Present address: MCD Biology, University of Colorado, Campus Box 347, Boulder, Colorado 80309, USA.

and, although their uteri are larger than at the neonatal stages, they remain smaller than those of wild-type siblings (Fig. 1d–f). Histological sections of the anterior region of the reproductive tract of *Wnt-7a*-mutant females show a fimbriated, ciliated epithelium typical of the proximal oviduct (Fig. 2a–d). More posterior sections show a less elaborately folded mucosa made up of a simple columnar epithelium, with a prominent layer of circularly arranged smooth muscle below (Fig. 2f). Although the mucosa is less well developed than in wild-type mice (Fig. 2e), this region resembles the isthmus region of the oviduct. Thus, regional differentiation occurs along the oviduct, but normal elongation and coiling of the duct do not occur.

In the uterus, the radial diameter is less than half that of wild-type siblings (Fig. 2g, h). The most prominent feature is a nearly complete absence of uterine glands, which are normally derived from the uterine epithelium, and a reduction in the mesenchymally derived uterine stroma, so that the transverse and circular muscles, which are present but reduced, lie closer to the endometrial epithelium (Fig. 2g, h). Thus, in the absence of *Wnt-7a*, both epithelial and mesenchymal differentiation are disrupted in the uterus. At both neonatal and adult stages, the ovaries, which are not

derived from the Müllerian duct, undergo normal follicular growth, ovulation and cycling in the adult (Fig. 2i, j). Thus, the normal hormonal regulation of reproductive development does not seem to be altered in *Wnt-7a* mutants, indicating that the phenotype we observe is unlikely to reflect an absence of the steroid hormones that coordinate uterine and ovarian development. As liveborn offspring were obtained from ovaries transplanted to recipient wild-type females (D. Rowitch and A.P.M., unpublished data), it is likely that a failure intrinsic to the Müllerian duct leads to female infertility.

Persistent Müllerian ducts are seen in homozygous *Wnt-7a*-mutant males throughout postnatal life. The non-regressed ducts appear as thin, undifferentiated tubes that run alongside the epididymis and the vas deferens from the testis to the urogenital sinus (Fig. 3a–d). The presence of a second duct seems to prevent the vas deferens from making a proper connection at its distal end (data not shown), a condition also seen in mice lacking Müllerian-inhibiting substance (MIS) or MIS receptor^{10,12}, resulting in a block to sperm passage. Analysis of the Müllerian ducts in adult males showed a thin, poorly differentiated structure consisting of little more than a uniform cuboidal epithelium with a thin layer of muscle beneath, and no regional differentiation along its length (Fig. 3e, f). The testes and Wolffian duct derivatives of *Wnt-7a*-mutant males appear normal and mature sperm fill the vas deferens (Fig. 3e, f). Thus, *Wnt-7a* is not involved in the development of the male reproductive system, but is required for Müllerian duct regression. The persistent Müllerian duct prevents the exit of sperm, leading to male sterility.

To determine how *Wnt-7a* may regulate Müllerian duct development, we analysed expression of *Wnt-7a* in the wild-type embryonic urogenital system. The Müllerian ducts arise from the coelomic epithelium in the mesonephric region of the mouse

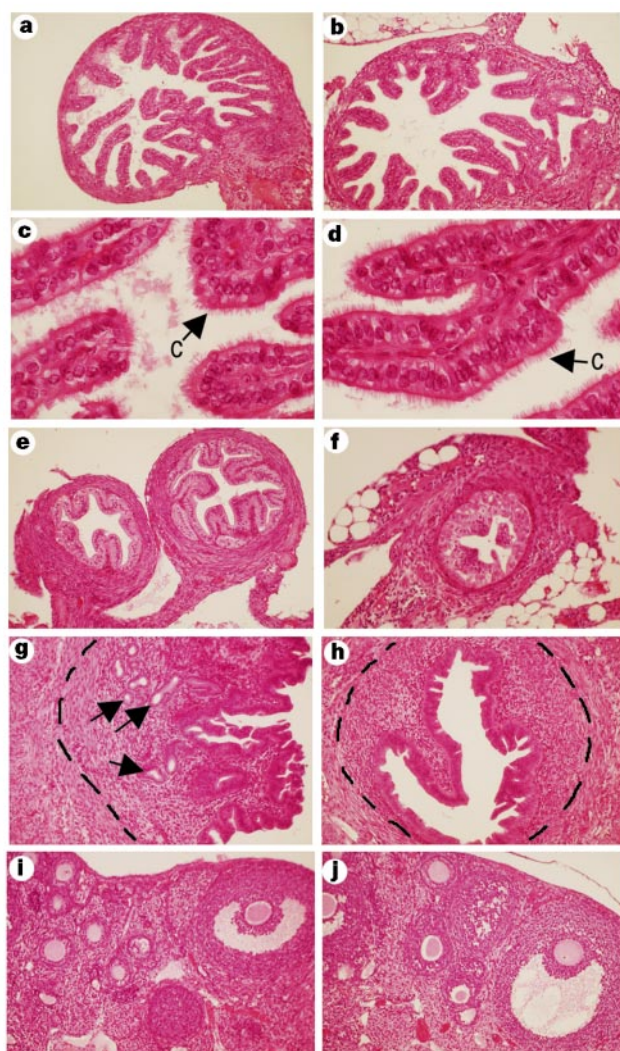


Figure 2 Sections through the reproductive tracts of adult wild-type and *Wnt-7a*-mutant mice. **a–f**, Oviducts of wild-type (**a, c, e**) and *Wnt-7a*-mutant (**b, d, f**) mice. **c**, cilia. **g, h**, Uteri of wild-type (**g**) and mutant (**h**) mice. The boundary between smooth muscle and stromal mesenchyme is indicated by dashed lines. Arrows in **g** indicate uterine glands, which are absent in the *Wnt-7a*-mutant. **i, j**, Ovaries of wild-type (**i**) and mutant (**j**) mice show cycling oocytes at comparable stages of follicular development.

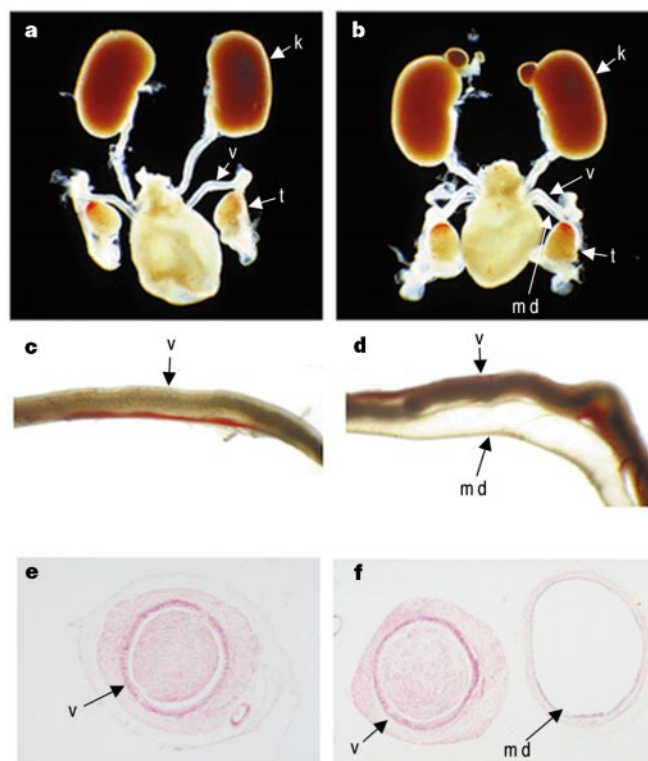


Figure 3 Male reproductive tracts in wild-type (**a, c, e**) and *Wnt-7a*-mutant (**b, d, f**) mice. **a, b**, Neonates. **c–f**, Adult stages seen in **c, d**, whole mounts and **e, f**, sections. *Wnt-7a*^{−/−} reproductive tracts look normal at both neonatal and adult stages except for the persistent Müllerian ducts seen in both whole mounts and sections. Sperm in the vas deferens of wild-type and mutant males is indicated by arrows. **k**, kidney; **v**, vas deferens; **md**, Müllerian duct; **t**, testis.

embryo, under the influence of Wnt-4, between 11.5 and 12.5 days *post coitum* (d.p.c.)¹⁶. *Wnt-7a* is expressed along the length of the Müllerian duct epithelium of both sexes from 12.5 to 14.5 d.p.c. (Fig. 4a, c and data not shown), but is then lost from the male epithelium, following Müllerian duct regression (data not shown). In the female, epithelial-specific *Wnt-7a* expression continues in the Müllerian duct derivatives throughout life, consistent with a function of Wnt-7a signalling in both fetal and adult ductal development¹⁷. To generate mutant *Wnt-7a*, we had inserted a neomycin-resistance gene into the second exon of the *Wnt-7a* locus¹⁴. This should generate a non-functional protein; however, truncated *Wnt-7a* transcripts can still be detected in *Wnt-7a*^{-/-} mice during embryogenesis. These transcripts are useful markers for the Müllerian duct epithelium. At 14.5 d.p.c., we still detected *Wnt-7a* transcripts in the Müllerian duct epithelium of both sexes of the mutant mice (Fig. 4b, d). Thus, by this criterion the epithelium is developing normally.

Our results indicate that the epithelial-derived *Wnt-7a* signal might function by regulating gene expression in the adjacent mesenchyme. To test this possibility, we analysed expression of the MIS type II receptor in the mesenchyme. Wild-type males (Fig. 4e) and females (Fig. 4g) express the receptor in mesenchymal cells surrounding the Müllerian duct at 14.5 d.p.c. Males also express the receptor in Sertoli cells of the testis⁵⁻⁷ (Fig. 4e). However, expression is absent from the Müllerian duct mesenchyme in *Wnt-7a*-mutant males (Fig. 4f) and females (Fig. 4h) but normal in the testis (Fig. 4f). Thus, *Wnt-7a* is required for regulation of MIS-receptor expression in the periductal mesenchyme and for subsequent regression of the Müllerian duct in males. These results indicate that *Wnt-7a* probably acts as a signal from the epithelium to the

surrounding mesenchyme during Müllerian duct development. The epithelial-to-mesenchymal signalling recalls the role of Wnt-7a in limb-bud patterning, where it signals from the dorsal ectoderm to the underlying mesenchyme to regulate expression of *Lmx-1b*, thereby patterning the dorsal limb^{14,15,18,19}.

Although the persistence of the Müllerian duct in *Wnt-7a*-mutant males is consistent with a loss of MIS signalling, mutants lacking either MIS or its receptor form well-developed oviducts and uteri in the male^{9,10,12}. In contrast, the persistent Müllerian duct in *Wnt-7a*^{-/-} males is a simple, undifferentiated tube with no regional organization. Thus, as in the female, *Wnt-7a* function is also necessary for further differentiation of the Müllerian duct. This indicates that Wnt-7a is not acting simply as an effector of MIS function in the Müllerian duct regression, but that it has a more important role in regulating morphogenesis of ductal derivatives.

Thus, our results indicate that signalling by Wnt-7a lies near the top of a differentiation pathway required for the development of the Müllerian duct in both sexes. After Wnt-4 initiates ductal formation¹⁶, Wnt-7a functions as an epithelial-to-mesenchymal signal, rendering the mesenchyme competent to respond to MIS signalling through the MIS receptor. However, this is probably just one aspect of a more general differentiation cascade regulated by Wnt-7a that is essential in the female for normal morphogenesis of Müllerian duct derivatives. Oestrogen and progesterone are important in development of the uterus after birth²⁰⁻²³. Here, their target is also the stromal mesenchyme. Whereas progesterone appears to function only in pregnancy²⁰, female mice with mutations in the oestrogen receptor show hypoplastic development of the uterine epithelium, stroma and muscle, and a marked loss of uterine glands in the adult²¹⁻²³. Thus, although the loss of *Wnt-7a* leads to a more severe phenotype than does the loss of the oestrogen receptor, the similarities between the two phenotypes indicates that *Wnt-7a* may be required for mediating a stromal response to oestrogen. □

Methods

Generation of *Wnt-7a*-mutant mice. The production of a likely null allele of *Wnt-7a* and the generation of *Wnt-7a*-mutant mice have been described¹⁴. The same restriction enzymes and probes described in ref. 14 were used for genotyping embryos and postnatal animals by Southern blotting here.

Histology. Reproductive tracts were dissected in phosphate-buffered saline (PBS) and fixed in Bouin's fixative. After dehydration through an ethanol series, the specimens were embedded in paraffin, sectioned at 7 µm, stained with haematoxylin and eosin, and photographed on a Leitz DM RB microscope using Ektachrome 64T film. Comparative photographs of wild-type and mutant siblings were taken at the same magnification.

In situ hybridization. *In situ* hybridizations to 6-µm paraffin sections were performed as described²⁴. Photographs were taken on a Leitz Aristoplan microscope using Fujichrome Velvia film, and blue and red filters were used to give a double-exposure image.

Whole-mount photography. Reproductive tracts from neonatal or adult mice were dissected in PBS. Whole-mount photographs of the reproductive tracts in PBS were taken with an Olympus SZH10 camera using Ektachrome 64T film.

Computer graphics. A Kodak RFS 2035 slide scanner was used to scan 35-mm slides into Adobe Photoshop 3.0 on a Power Macintosh 8100. Composite figures were organized and labelled in Photoshop.

Received 27 April; accepted 10 August 1998.

1. Jossio, N. *et al.* Anti-Müllerian hormone, the Jost factor. *Recent Prog. Horm. Res.* **48**, 1-59 (1993).
2. Cate, R. L. *et al.* Isolation of the bovine and human genes for Müllerian-inhibiting substance and expression of the human gene in animal cells. *Cell* **45**, 685-698 (1986).
3. Picard, J.-Y., Benarous, R., Guerrier, D., Jossio, N. & Kahn, A. Cloning and expression of cDNA for anti-Müllerian hormone. *Proc. Natl Acad. Sci. USA* **83**, 5464-5468 (1986).
4. Jost, A. Problems of fetal endocrinology: the gonadal and hypophyseal hormones. *Recent Prog. Horm. Res.* **8**, 379-418 (1953).
5. Di Clemente, N. *et al.* Cloning, expression and alternative splicing of the receptor for anti-Müllerian hormone. *Mol. Endocrinol.* **8**, 1006-1020 (1994).
6. Baarends, W. M. *et al.* A novel member of the transmembrane serine/threonine kinase receptor family is specifically expressed in the gonads and in mesenchymal cells adjacent to the Müllerian duct. *Development* **120**, 189-197 (1994).
7. Teixeira, J. *et al.* Developmental expression of a candidate Müllerian inhibiting substance type II receptor. *Endocrinology* **137**, 160-165 (1996).

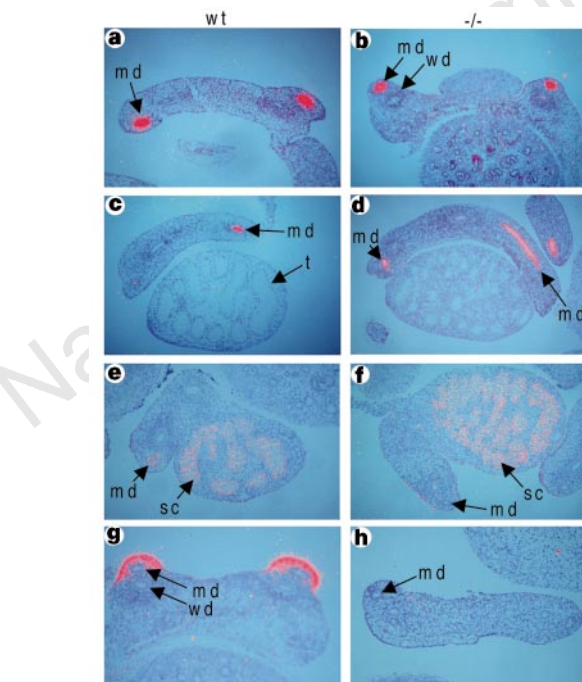


Figure 4 *In situ* hybridization was used to analyse *Wnt-7a* and Müllerian-inhibiting substance (MIS)-receptor messenger RNA expression in the Müllerian duct at 14.5 d.p.c. **a, b**, *Wnt-7a* transcripts are present in the Müllerian duct (md) epithelium of **a**, wild-type and **b**, *Wnt-7a*-mutant female embryos. wd, Wolffian duct. **c, d**, Similarly, *Wnt-7a* mRNA is detected in the Müllerian ducts of **c**, wild-type and **d**, *Wnt-7a*-mutant males. **e**, In wild-type males, MIS receptor mRNA is detected in the mesenchyme surrounding the Müllerian duct and in testicular Sertoli cells (sc). **g**, MIS receptor expression in wild-type females is polarized, as it is not present in mesenchyme between the Müllerian and Wolffian ducts. **f, h**, MIS-receptor expression in Müllerian duct mesenchyme is not detectable in *Wnt-7a*-mutant males (**f**) or females (**h**), although it persists in the Sertoli cells. t, testis.

8. Knebelmann, B. *et al.* Anti-Müllerian hormone Bruxelles: a non-sense mutation associated with the persistent Müllerian duct syndrome. *Proc. Natl. Acad. Sci. USA* **88**, 3767–3771 (1991).
9. Jossio, N. *et al.* Clinical aspects and molecular genetics of the persistent Müllerian duct syndrome. *Clin. Endocrinol.* **47**, 137–144 (1997).
10. Behringer, R. B., Finegold, M. J. & Cate, R. L. Müllerian-inhibiting substance function during mammalian sexual development. *Cell* **79**, 415–425 (1994).
11. Imbeaud, S. *et al.* A 27 base-pair deletion of the anti-Müllerian type II receptor gene is the most common cause of the persistent Müllerian duct syndrome. *Hum. Mol. Genet.* **5**, 1269–1277 (1996).
12. Mishina, Y. *et al.* Genetic analysis of the Müllerian-inhibiting substance signal transduction pathway in mammalian sexual differentiation. *Genes Dev.* **10**, 2577–2587 (1996).
13. Cadigan, K. M. & Nusse, R. Wnt signaling: a common theme in animal development. *Genes Dev.* **11**, 3286–3305 (1997).
14. Parr, B. A. & McMahon, A. P. Dorsalizing signal Wnt-7a required for normal polarity of D-V and A-P axes of mouse limb. *Nature* **374**, 350–353 (1995).
15. Cygan, J. A., Johnson, R. L. & McMahon, A. P. Novel regulatory interactions revealed by studies of murine limb pattern in Wnt-7a and En-1 mutants. *Development* **124**, 5021–5032 (1997).
16. Vainio, S., Heikkilä, M., Kispert, A., Chin, N. & McMahon, A. P. Female development in mammals regulated by Wnt-4 signaling. *Nature* (submitted).
17. Miller, C. & Sassoon, D. Wnt-7a maintains appropriate uterine patterning during the development of the mouse female reproductive tract. *Development* **125**, 3201–3211 (1998).
18. Riddle, R. D. *et al.* Induction of the LIM homeobox gene *Lmx1* by WNT7A establishes dorsoventral pattern in the vertebrate limb. *Cell* **83**, 631–640 (1995).
19. Vogel, A., Rodriguez, C., Warnken, W. & Izpisua-Belmonte, J. C. Dorsal cell fate specified by chick *Lmx1* in the vertebrate limb. *Nature* **378**, 716–720 (1995).
20. Lydon, J. P. *et al.* Mice lacking progesterone receptor exhibit pleiotropic reproductive abnormalities. *Genes Dev.* **9**, 2266–2278 (1995).
21. Lubahn, D. B. *et al.* Alteration of reproductive function but not prenatal sexual development after insertional disruption of the mouse estrogen receptor gene. *Proc. Natl. Acad. Sci. USA* **90**, 11162–11166 (1993).
22. Korach, K. S. Insights from the study of animals lacking functional estrogen receptors. *Science* **266**, 1524–1527 (1994).
23. Korach, K. S. *et al.* Estrogen receptor gene disruption: molecular characterization and experimental and clinical phenotypes. *Recent Prog. Horm. Res.* **51**, 159–187 (1996).
24. Parr, B. A., Shea, M. J., Vassileva, G. & McMahon, A. P. Mouse Wnt genes exhibit discrete domains of expression in the early embryonic CNS and limb buds. *Development* **119**, 247–261 (1993).

Acknowledgements. We thank B. Klumpp for histology and W. Baarends for the MIS type II receptor probe. This work was supported by a grant from the NIH (to A.P.M.).

Correspondence and requests for materials should be addressed to A.P.M. (e-mail: amcmahon@biosun.harvard.edu).

Spanning binding sites on allosteric proteins with polymer-linked ligand dimers

Richard H. Kramer* & Jeffrey W. Karpen†

* Department of Molecular and Cellular Pharmacology, University of Miami School of Medicine, Miami, Florida 33101, USA

† Department of Physiology and Biophysics, University of Colorado School of Medicine, Denver, Colorado 80262, USA

One approach to drug design involves determination of the structure of binding sites on target proteins to provide templates for ligand construction. Alternatively, random combinations of chemical groups can be used to generate diverse molecules for screening in the search for effective compounds¹. Here we report a strategy for developing potent ligands for proteins with multiple binding sites, which combines elements of both approaches: 'polymer-linked ligand dimers', in which two ligands are joined by a polymer chain of variable length. We find that polymer-linked ligand dimers containing two cyclic GMP moieties are up to a thousand times more potent than cyclic GMP in activating cyclic-nucleotide-gated channels and cGMP-dependent protein kinase. Each target protein responds optimally to a polymer-linked ligand dimer with a different average polymer length, even though their cyclic-nucleotide-binding sites are conserved. The tuning of polymer-linked ligand dimers indicates that each protein has a unique spacing of binding sites and provides an estimate of the distance between these sites. As optimal ligands are selected empirically, the polymer-linked ligand dimer strategy enables potent and selective agents to be identified without requiring previous structural information about the target proteins.

The activities of cyclic nucleotides are mediated by allosteric effector proteins such as the cyclic-nucleotide-gated (CNG) channels in vertebrate photoreceptors and olfactory neurons^{2,3}

and the protein kinases activated by cGMP (PKG; ref. 4) or cAMP (PKA; ref. 5). Each of these proteins contains four cyclic-nucleotide-binding sites, either on individual subunits in CNG channels⁶, or grouped two to a subunit as in PKA and PKG^{4,5}. Although their binding sites share considerable homology⁷, these target proteins differ in their apparent affinities for cyclic-nucleotide derivatives, enabling agonists or antagonists to be identified^{8–12} that preferentially affect specific proteins, albeit with a low apparent affinity.

We synthesized polymer-linked ligand dimers (PLDs) by reacting a sulphhydryl derivative of cGMP¹³ with a bifunctional vinylsulphone-derivatized polyethylene glycol (PEG). This produced a barbell-shaped molecule containing two cGMPs, connected to both ends of a flexible polymer through a thioether linkage at the 8-position of the purine (Fig. 1a). Conjugation of bulky groups at the 8-position, rather than being detrimental, increases the apparent affinity of ligands for CNG channels^{13–17}. PLDs of different lengths were synthesized and named according to the relative molecular mass of the polymer and the nature of the ligand. Thus, a PLD composed of a 3,400 dalton PEG conjugated to two cGMPs is referred to as 3400PEG-(cGMP)₂. We estimated the average (r.m.s.) lengths of each PLD from previous determinations of the length of specific PEGs¹⁸, assuming that polymer length is proportional to the square root of the number of monomers¹⁹.

How does changing the polymer length of a PLD affect its ability to bind a protein with multiple ligand-binding sites? Figure 1b shows the interaction between a CNG channel and three distinct PLDs. Each CNG channel subunit contains a cytoplasmic cyclic-nucleotide-binding site. When the PLD is much shorter than the distance between two binding sites, the apparent affinity of the PLD should be about twice that of cGMP, assuming that the polymer does not affect the ability of the cGMP moieties to bind or activate the channel.

When the average length of the PLD matches the distance between the binding sites on the channel, the apparent affinity should increase dramatically. After one cGMP moiety binds, the diffusion of its tethered cognate cGMP will be constrained within a hemisphere of radius equal to the r.m.s. length of the PLD. For a radius of 39 Å, the effective concentration (C_{eff}) will be 13.4 mM, which is much higher than the apparent K_d for cGMP (for example, 4 μ M). Thus, once one cGMP has bound, the probability of binding of its cognate cGMP will increase, enhancing the overall affinity. For larger PLDs, the C_{eff} of the cognate ligand should decline, reducing the apparent affinity.

We tested this prediction by applying PLDs, as well as a control polymer-linked ligand monomer (PLM) containing only one cGMP moiety conjugated to a 5,000 dalton PEG, onto excised patches containing rat olfactory²⁰ (OLF), or bovine rod photoreceptor²¹ (RET) CNG channels. Dose-response curves were fitted with the Hill equation to determine the apparent affinity of the channels for each PLD and the PLM. For the olfactory channel (Fig. 2a), the apparent affinity of the PLM was slightly higher than that of cGMP ($K_{1/2}$ was 1.4 and 3.1 μ M, respectively), indicating that conjugation of PEG onto the 8-position of cGMP slightly improved activity. As expected, the apparent affinity of the shortest PLD, 282PEG-(cGMP)₂, was slightly higher ($K_{1/2}$ = 0.85 μ M) than that shown by cGMP or the PLM. The Hill coefficient, indicative of the minimum number of individual molecules required for significant activation, was similar for cGMP, the PLM, and 282PEG-(cGMP)₂ (1.7, 2.0 and 2.0, respectively).

In contrast to the moderate increase in affinity shown by 282PEG-(cGMP)₂, 3400PEG-(cGMP)₂, a larger PLD, had a dramatic effect: $K_{1/2}$ for this PLD was 12 nM, which is 260-fold lower than for cGMP. An even larger PLD, 20000PEG-(cGMP)₂, gave a less marked increase in the apparent affinity ($K_{1/2}$ = 250 nM). The Hill coefficient values for 3400PEG-(cGMP)₂ and 20000PEG-(cGMP)₂ were 1.1 and 0.8, respectively, about half as large as for the PLM or for the smallest PLD, 282PEG-(cGMP)₂. This is consistent with the larger, but not the smallest PLDs simultaneously occupying two binding sites on the channel.

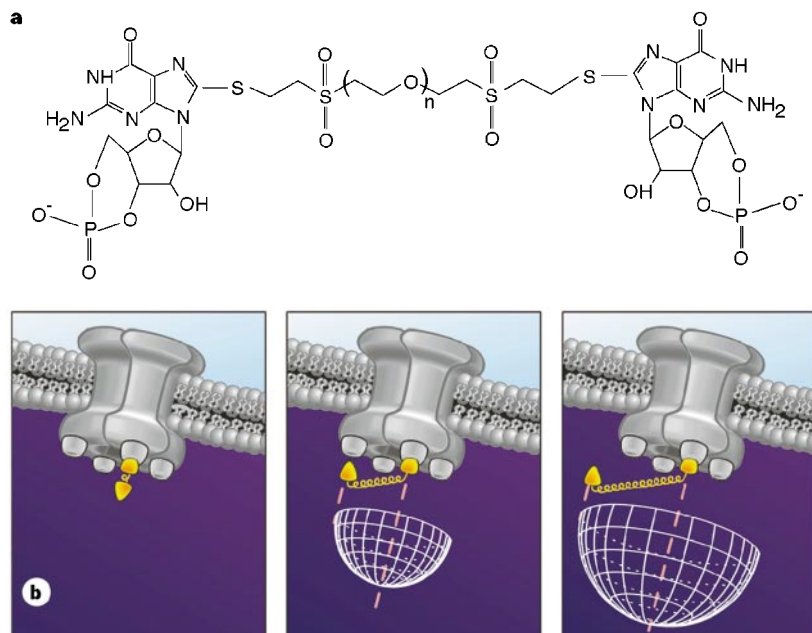


Figure 1 Polymer-linked ligand dimers (PLDs). **a**, Structure of a PLD containing two cGMP moieties and a poly(ethylene glycol) (PEG) linker, with n ethylene glycol units. **b**, Diagrams of PLDs binding to a channel with four ligand-binding sites: PLDs are shown with polymers whose average length is too short (left), just right (centre), or longer (right) than necessary, to allow the ligands to span two binding sites on the channel. In the optimal PLD, following binding of one ligand, the effective concentration (C_{eff}) of the tethered unbound ligand is calculated from the volume of the hemisphere with radius r (in cm) circumscribed by the tethered molecule. Thus $C_{\text{eff}} = 1,000/(N_A 2/3\pi r^3)$, where N_A is Avogadro's number.

The effect of PLDs on rod photoreceptor CNG channels was similar, with the intermediate PLD, 3400PEG-(cGMP)₂, having a much higher apparent affinity and a lower Hill coefficient than cGMP or the shortest PLD (Fig. 2b). However, unlike their action on olfactory CNG channels, all PLDs were only partial agonists of photoreceptor CNG channels, giving <75% maximal activation compared with cGMP. Conversely, the PLM was a full agonist. The difference between the PLDs and the PLM indicates that the polymer chain spanning two sites interferes slightly with gating or ion permeation through the photoreceptor channel.

Results from kinetic experiments support the idea that PLDs simultaneously occupy two binding sites. The dissociation rate of a PLD simultaneously bound to two sites, $k_{\text{off(PLD)}}$, is given as

$$k_{\text{off(PLD)}} = 2k_{\text{off}}K_d/(C_{\text{eff}} + K_d) \quad (1)$$

where k_{off} is the intrinsic 'off' rate of cGMP from a single binding site, C_{eff} is the effective concentration of cGMP at the second site, and K_d is the equilibrium dissociation constant of cGMP. Equation (1) assumes that $k_{\text{off(PLD)}}$ is equal to twice the k_{off} times the probability that the second site is not occupied. In the case of the dissociation of 2000PEG-(cGMP)₂ from the olfactory channel, C_{eff} is predicted to be 13.4 mM and the apparent K_d for cGMP is 4 μM , yielding a PLD dissociation rate that is 3400-fold slower than for cGMP. As predicted, dissociation of this PLD was much slower (Fig. 3a) than that of either cGMP or the shortest PLD, 282PEG-(cGMP)₂, so channels remained open for minutes, even with continuous superfusion with agonist-free solution. When a cGMP dissociates, it therefore has a high probability of rebinding before its partner can also dissociate. The 'off' rate of a longer PLD, 20000PEG-(cGMP)₂, was slowed less dramatically, consistent with the lower predicted C_{eff} . Thus, the decrease in dissociation rate most probably accounts for the enhanced apparent affinity shown by PLDs for the olfactory channel.

The dissociation of a ligand from a single binding site is normally unaffected by the concentration of free ligand in solution. However, if PLDs bind to two sites at once, occupancy of either site with a competing ligand should speed its dissociation: Fig. 3b shows that when channels are first loaded with a PLD, subsequent application of cGMP does indeed accelerate the 'off' rate. A high concentration (<1 mM) is required, however, because free cGMP must compete with PLD ligands that have effective concentrations of several millimolar. Our observation that addition of free ligand (cGMP) increases the dissociation rate of a bound ligand (the PLD) provides

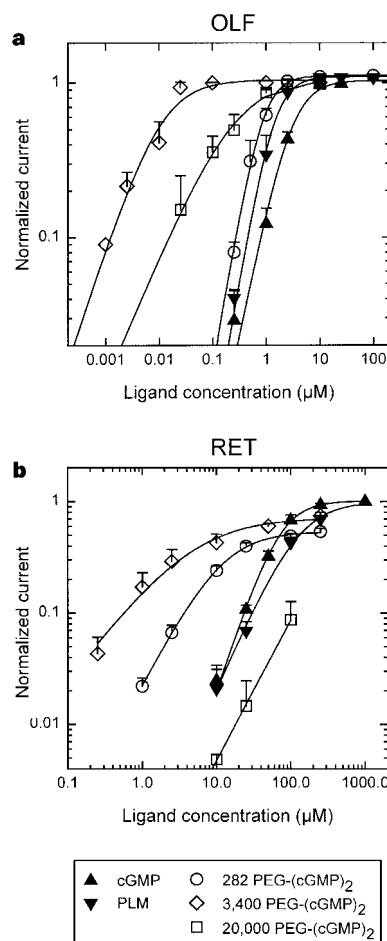


Figure 2 Activation of cyclic-nucleotide-gated (CNG) channels by polymer-linked ligand dimers (PLDs) and monomeric cyclic nucleotides. **a**, **b**, Dose-response curves of activation of olfactory (OLF) (**a**) and rod photoreceptor (RET) (**b**) CNG channels by cGMP, three different PLDs and the control polymer-linked ligand monomer (PLM). Responses from each patch were normalized to the response elicited by 2 mM cGMP. Fits to the Hill equation yield $K_{1/2}$ and n values for the RET channels as follows: cGMP: $K_{1/2} = 72 \mu\text{M}$, $n = 2.0$; 282PEG-(cGMP)₂: $K_{1/2} = 11 \mu\text{M}$, $n = 1.3$; 3400PEG-(cGMP)₂: $K_{1/2} = 4.7 \mu\text{M}$, $n = 0.9$; 20000PEG-(cGMP)₂: $K_{1/2} > 75 \mu\text{M}$. $K_{1/2}$ and n values for OLF channels are given in the text.

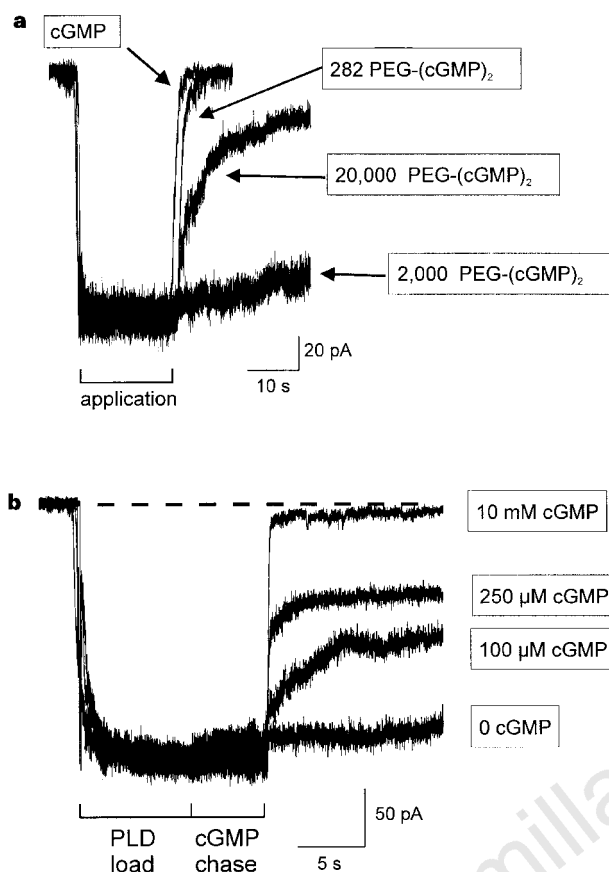


Figure 3 Dissociation of polymer-linked ligand dimers (PLDs) from cyclic-nucleotide-gated (CNG) channels. **a**, Tail currents reflecting dissociation of cGMP and various PLDs. Free agonists were applied for the indicated period of time and washed away with continuous perfusion of saline. **b**, Effect of free cGMP on the dissociation of bound PLDs. CNG channels were loaded with 2000PEG-(cGMP)₂ for 15 s. Free PLD was removed during the chase period and replaced with various concentrations of cGMP. The patch was then perfused with saline to examine the extent of dissociation of the PLDs. Kinetic experiments in **a** and **b** were done with OLF channels.

strong evidence that the PLD simultaneously occupies two sites on the channel.

All CNG channels have four binding sites, but in other respects their molecular dimensions may differ, including the distance between binding sites. To determine this distance, we used a series of PLDs containing PEG polymers with relative molecular masses between 282 and 20,000, corresponding to r.m.s. lengths of 15–123 Å. Figure 4a shows that the optimal PLD for activating rod photoreceptor CNG channels (1200PEG-(cGMP)₂; r.m.s. length, 30 Å) is shorter than that for activating olfactory channels (2000PEG-(cGMP)₂; r.m.s. length, 39 Å). These findings indicate that there might be a wider spacing of binding sites in the olfactory channel, consistent with a previous report that the olfactory channel has a wider pore than the photoreceptor channel²². However, it is difficult to determine exact distances because of the flexibility, polydispersity and finite number of PLDs used (see Supplementary Information). Rather than increasing abruptly at the optimal length, the apparent affinity of PLDs starts to increase before the optimal length is reached. This probably results from elasticity of the PLD and possibly of the CNG channel, giving a shorter than optimal PLD some finite probability of spanning two binding sites. Using polymers that are stiffer than PEG, it may be possible to generate PLDs with little elasticity which are thus more finely tuned to specific proteins.

We also examined the effect of the PLDs on PKG. Each PKG

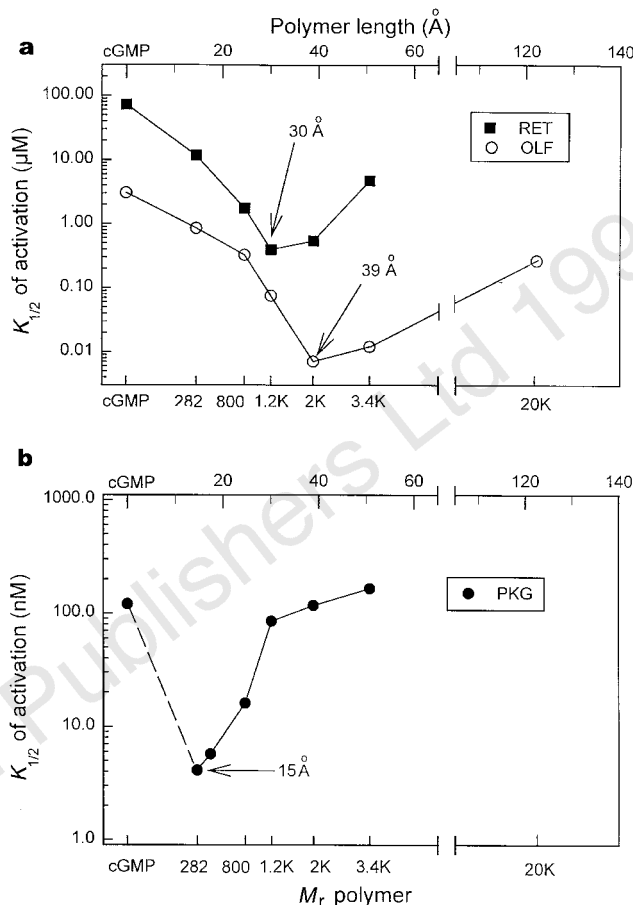


Figure 4 Identification of optimal polymer-linked ligand dimers (PLDs) for cGMP-binding proteins. **a**, Activation of CNG channels (RET and OLF). **b**, Activation of cGMP-dependent protein kinase (PKG). Arrows indicate PLDs optimal for activating each target protein. Polymer lengths were estimated from previous determinations¹⁸, assuming an increase with the square root of the M_r (ref. 19).

subunit has a regulatory domain containing one high- and one low-affinity cGMP-binding site. This domain shares extensive homology with the regulatory subunit of PKA⁷, in which two sites are separated by 26 Å, as indicated by X-ray crystallography²³. Thus, the two heterologous sites within a PKG monomer may be 26 Å apart, whereas the distance between homologous sites across two subunits is unknown. Figure 4b shows that the optimal PLD for PKG was 282PEG-(cGMP)₂. This PLD activated PKG with an apparent affinity that was 30-fold higher than for cGMP ($K_{1/2} = 4$ nM for the PLD; $K_{1/2} = 120$ nM for cGMP) and 275-fold higher than for the PLM ($K_{1/2} = 1.1$ μM). The increase in apparent affinity of 282PEG-(cGMP)₂ indicates that it probably simultaneously occupies two sites on PKG; however, the short r.m.s. length indicates that PKG activation by this PLD primarily involves binding between subunits to homologous sites <20 Å apart, rather than binding within a single PKG subunit to heterologous sites.

For all three cyclic-nucleotide-activated proteins (photoreceptor and olfactory channels, and PKG), at least some PLDs had apparent affinities that were much higher than cGMP. In fact, specific PLDs were nearly one thousand times more effective at activating CNG channels, making these molecules the most potent agonists identified so far. The optimal PLD was different for each target protein, with the 282-, 1200- and 2000PEG-(cGMP)₂ molecules being optimal for PKG and the photoreceptor and olfactory channels, respectively. Hence, our PLD strategy should reveal agents with a

selectivity for specific proteins, as we have shown for 2000PEG-(cGMP)₂ (Fig. 4) which, to our knowledge, is the first agonist to activate a CNG channel (olfactory) more potently than PKG. At least some of the PLDs are membrane-permeant (see Supplementary Information), which will enable functional roles of PKG and of CNG channels to be dissected in intact cells. Given the appropriate ligand composition and polymer size, highly effective antagonists of individual cyclic-nucleotide-binding proteins can be designed.

The PLD approach has several remarkable features. Selective ligands can be produced for proteins with similar or identical ligand-binding sites, without previous structural information about the target proteins. The high potency and selectivity of PLDs raises the possibility that this approach to ligand design can be applied to many proteins with multiple binding sites, including ligand-gated ion channels, transport proteins and allosteric enzymes. □

Methods

Synthesis of PLDs. PLDs were synthesized by reacting 8-thio cGMP¹³ with bifunctional PEG vinylsulphone (VS-PEG-VS) (Shearwater Polymers) at 37 °C and pH 7.5 for 24–48 h, and purified by reverse-phase HPLC using a preparative C18 column eluted with a methanol gradient. Reactions were run either with an excess of 8-thio cGMP to bias formation of PLDs (cGMP-PEG-cGMP), or with an excess of VS-PEG-VS to bias formation of half-reacted monomers (cGMP-PEG-VS). To distinguish between these products, their elution profiles were compared at 274 nm, the peak absorbance wavelength of thio-substituted cyclic nucleotides. The composition of 282 PEG(cGMP)₂ was confirmed by mass spectrometry. The concentration of each PLD was determined by measuring its absorbance, assuming an extinction coefficient of 35,400 M⁻¹ cm⁻¹ (17,700 M⁻¹ cm⁻¹ for the PLM)¹³. A capped PEG (M_r 5,000) with a single vinylsulphone group (methoxy-PEG-VS) was used to produce a control PLM.

Expression and recording of CNG channel activity. Homomeric channels containing α-subunits of the bovine rod²¹ or the rat olfactory CNG channel²⁰ were expressed in *Xenopus* oocytes. Whereas native CNG channels contain α- and β-subunits^{24–27}, homotetrameric channels composed solely of α-subunits can be expressed, and have properties in common with native channels³. Patch pipettes containing standard saline (115 mM NaCl, 1 mM EDTA, 5 mM EGTA, 5 mM HEPES; pH 7.5) were used to obtain excised inside-out patches. Membrane potential was held constant at –75 mV in all experiments. cGMP and derivatives in standard saline were superfused over excised patches at 21–23 °C. All experimental results were from at least three patches. Variability among data is expressed as mean ± standard error of the mean.

Measurement of PKG activity. PKG activity was assayed by determining incorporation of ³²P from [γ-³²P]ATP into a heptapeptide substrate²⁸. Bovine recombinant PKG type 1α was obtained from Calbiochem.

Received 26 May; accepted 23 July 1998.

- Gallop, M. A., Barrett, R. W., Dower, W. J., Fodor, S. P. & Gordon, E. M. Applications of combinatorial technologies to drug discovery. 1. Background and peptide combinatorial libraries. *J. Med. Chem.* **37**, 1233–1251 (1994).
- Finn, J. T., Grunwald, M. E. & Yau, K.-W. Cyclic nucleotide-gated ion channels: an extended family with diverse functions. *Annu. Rev. Physiol.* **58**, 395–426 (1996).
- Zagotta, W. N. & Siegelbaum, S. A. Structure and function of cyclic nucleotide-gated channels. *Annu. Rev. Neurosci.* **19**, 235–263 (1996).
- Francis, S. H. & Corbin, J. D. Progress in understanding the mechanism and function of cyclic GMP-dependent protein kinase. *Adv. Pharm.* **26**, 115–170 (1994).
- Taylor, S. S., Buechler, J. A. & Yonemoto, W. cAMP-dependent protein kinase: framework for a diverse family of regulatory enzymes. *Annu. Rev. Biochem.* **59**, 971–1005 (1990).
- Liu, D. T., Tibbs, G. R. & Siegelbaum, S. A. Subunit stoichiometry of cyclic nucleotide-gated channels and effects of subunit order on channel function. *Neuron* **16**, 983–990 (1996).
- Shabb, J. B. & Corbin, J. D. Cyclic nucleotide-binding domains in proteins having diverse functions. *J. Biol. Chem.* **267**, 5723–5726 (1992).
- Botelho, L. H., Rothermel, J. D., Coombs, R. V. & Jastorff, B. cAMP analog antagonists of cAMP action. *Meth. Enzymol.* **159**, 159–172 (1988).
- Butt, E., Van Bemmelen, M., Fischer, L., Walter, U. & Jastorff, B. Inhibition of cGMP-dependent protein kinase by (Rp)-guanosine 3',5'-monophosphorothioates. *FEBS Lett.* **263**, 47–50 (1990).
- Butt, E., Eigenthaler, M. & Genieser, H. G. (Rp)-8-pCPT-cGMPs, a novel cGMP-dependent protein kinase inhibitor. *Eur. J. Pharmacol.* **269**, 265–268 (1994).
- Kramer, R. H. & Tibbs, G. R. Antagonists of cyclic nucleotide-gated channels and molecular mapping of their site of action. *J. Neurosci.* **16**, 1285–1293 (1996).
- Wei, J.-Y., Cohen, E. D., Yan, Y.-Y., Genieser, H.-G. & Barnstable, C. J. Identification of competitive antagonists of the rod photoreceptor cGMP-gated cation channel: beta-phenyl-1,N2-etheno-substituted cGMP analogues as probes of the cGMP-binding site. *Biochemistry* **35**, 16815–16823 (1996).
- Brown, R. L., Bert, R. J., Evans, F. E. & Karpen, J. W. Activation of retinal rod cGMP-gated channels: what makes for an effective 8-substituted derivative of cGMP? *Biochemistry* **32**, 10089–10095 (1993).

- Zimmerman, A. L., Yamanaka, G., Eckstein, F., Baylor, D. A. & Stryer, L. Interaction of hydrolysis-resistant analogs of cyclic GMP with the phosphodiesterase and light-sensitive channel of retinal rod outer segments. *Proc. Natl Acad. Sci. USA* **82**, 8813–8817 (1985).
- Cartetta, A., Cavagioni, A. & Sorbi, R. T. Binding stoichiometry of a fluorescent cGMP analogue to membranes of retinal rod outer segments. *Eur. J. Biochem.* **153**, 49–53 (1985).
- Koch, K.-W. & Kaupp, U. B. Cyclic GMP directly regulates a cation conductance in membranes of bovine rods by a cooperative mechanism. *J. Biol. Chem.* **260**, 6788–6800 (1985).
- Tanaka, J. C., Eccleston, J. F. & Furman, R. E. Photoreceptor channel activation by nucleotide derivatives. *Biochemistry* **28**, 2776–2784 (1989).
- Knoll, D. & Hermans, J. Polymer–protein interactions. Comparison of experiment and excluded volume theory. *J. Biol. Chem.* **258**, 5710–5715 (1983).
- Florey, P. J. *Principles of Polymer Chemistry* (Cornell Univ. Press, Ithaca, New York, 1953).
- Dhallan, R. S., Yau, K.-W., Schrader, K. A. & Reed, R. R. Primary structure and functional expression of a cyclic-nucleotide-activated channel from olfactory neurons. *Nature* **347**, 184–187 (1990).
- Kaupp, U. B. *et al.* Primary structure and functional expression from complementary DNA of the rod photoreceptor cyclic GMP-gated channel. *Nature* **342**, 762–766 (1989).
- Goulding, E. H., Tibbs, G. R., Liu, D. & Siegelbaum, S. A. Role of H5 domain in determining pore diameter and ion permeation through cyclic nucleotide-gated channels. *Nature* **364**, 61–64 (1993).
- Su, Y. *et al.* Regulatory subunit of protein kinase A: structure of deletion mutant with cAMP binding domains. *Science* **269**, 807–813 (1995).
- Chen, T.-Y. *et al.* A new subunit of the cyclic nucleotide-gated cation channel in retinal rods. *Nature* **362**, 764–767 (1993).
- Liman, E. R. & Buck, L. B. A second subunit of the olfactory cyclic nucleotide-gated channel confers high sensitivity to cAMP. *Neuron* **13**, 611–621 (1994).
- Bradley, J., Li, J., Davidson, N., Lester, H. A. & Zinn, K. Heteromeric olfactory cyclic nucleotide-gated channels: a subunit that confers increased sensitivity to cAMP. *Proc. Natl Acad. Sci. USA* **91**, 8890–8894 (1994).
- Körschen, H. G. *et al.* A 240 kDa protein represents the complete beta subunit of the cyclic nucleotide-gated channel from rod photoreceptor. *Neuron* **15**, 627–636 (1995).
- Sekhar, K. R. *et al.* Relaxation of pig coronary arteries by new and potent cGMP analogs that selectively activate type I alpha, compared with type I beta, cGMP-dependent protein kinase. *Mol. Pharmacol.* **42**, 103–108 (1992).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements. We thank P. Fennessey and L. Miller for mass spectrometry, E. Nelson for technical help, and C. Barski for assistance with artwork. This work was supported by grants from the NIH (to R.H.K. and J.W.K.) and the American Heart Association.

Correspondence and requests for materials should be addressed to R.H.K. (e-mail: rkramer@chroma.med.miami.edu) or to J.W.K. (e-mail: Jeffrey.karpen@uchsc.edu).

Bmk1/Erk5 is required for cell proliferation induced by epidermal growth factor

Yutaka Kato, Richard I. Tapping, Shuang Huang, Mark H. Watson, Richard J. Ulevitch & Jiing-Dwan Lee

Department of Immunology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

Epidermal growth factor (EGF) induces cell proliferation in a variety of cell types by binding to a prototype transmembrane tyrosine kinase receptor^{1,2}. Ligation of this receptor by EGF activates Erk1 and Erk2, members of the mitogen-activated protein (MAP) kinase family, through a Ras-dependent signal transduction pathway^{3–5}. Despite our detailed understanding of these events, the exact mechanism by which EGF causes cells to proliferate is unclear. Big MAP kinase (Bmk1), also known as Erk5, is a member of the MAP kinase family that is activated in cells in response to oxidative stress, hyperosmolarity and treatment with serum^{6,7}. Here we show that EGF is a potent activator of Bmk1. In contrast to Erk1/2, EGF-mediated activation of Bmk1 occurs independently of Ras and requires the MAP-kinase kinase Mek5. Expression of a dominant-negative form of Bmk1 blocks EGF-induced cell proliferation and prevents cells from entering the S phase of the cell cycle. These results demonstrate that Bmk1 is part of a distinct MAP-kinase signalling pathway that is required for EGF-induced cell proliferation and progression through the cell cycle.

We previously reported that Bmk1 is activated in cells in response to serum and that it stimulates early gene expression through phosphorylation of the transcription factor MEF2C⁷. In studies to identify serum component(s) responsible for Bmk1 activation, we discovered that whereas other typical MAP kinase inducers such as

phorbol myristate acetate (PMA), tumour-necrosis factor (TNF) and interleukin-1 (IL-1) had no effect, EGF was a potent activator of Bmk1 (Fig. 1a). The activation of Bmk1 was detectable within 5 min of EGF addition and was maximal within 15 min (Fig. 1b). In HeLa cells, Bmk1 was activated with as little as 0.1 ng ml⁻¹ (16 pM) of EGF and 1 ng ml⁻¹ of EGF activated Bmk1 maximally (Fig. 1c).

The EGF receptor requires its tyrosine kinase activity in order to signal and mediate cellular mitogenic responses². To determine whether the tyrosine kinase activity of the EGF receptor is required for EGF-mediated Bmk1 activation, we used compound 32, a highly specific inhibitor of EGF-receptor tyrosine kinase activity⁸. Compound 32 treatment prevented EGF-mediated activation of both Bmk1 and ERKs (Fig. 1d). In contrast, compound 32 had no effect on the cellular activation of Bmk1 and ERKs in response to hyperosmolar conditions.

Ras is an essential element for EGF-induced activation of both ERK and JNK MAP kinases^{9,10}. Expression of dominant active forms of Ras (Ras-G12V), or its effector Raf-1 (Raf/BXB), significantly activated endogenous ERK, but failed to activate endogenous Bmk1 in HeLa cells (Fig. 2a). In contrast, the expression of a constitutively active form of Mek5 (Mek5(D)) activated Bmk1 without affecting the activity of ERK. Moreover, when we expressed a dominant-negative form of Ras (Ras-T17N) in HeLa cells, EGF-induced ERK activation was blocked, whereas EGF-mediated Bmk1 activation was unaffected (Fig. 2b).

Mek5 is known to activate Bmk1 directly^{7,11}. We have found that Mek5 is activated in cells after EGF treatment, the time course of activation being indistinguishable from that of EGF-mediated activation of Bmk1 (Fig. 2c). Moreover, adenoviral gene delivery and expression of a dominant-negative form of Mek5, Mek5(A), inhibited EGF-induced stimulation of Bmk1 (Fig. 2d). This effect required Mek5(A) expression because EGF-induced activation of Bmk1 was unaffected in mock or control adenovirus-infected cells. In addition, the inhibitory effect of Mek5(A) seemed to be specific for Bmk1 activity as Mek5(A) expression did not modify the EGF-induced activation of ERK.

As Bmk1 appears to be a cellular mediator of EGF-induced signalling, we investigated the involvement of Bmk1 in EGF-induced cell proliferation using MCF10A cells, an immortalized breast epithelial cell line requiring EGF for growth¹². To this end, a

dominant-negative form of Bmk1, Bmk1(AEF)⁷, was expressed in MCF10A cells using adenoviral gene transduction. As the classic ERK pathway has been implicated in EGF-induced signalling, dominant-negative forms of Erk2, Erk2(AEF)⁷, or Mek1, Mek1(A)⁷, were also expressed in MCF10A cells. Protein expression of dominant-negative kinases was determined by western blotting (see Fig. 3b for dominant-negative Erk2 and Bmk1; data not shown for dominant-negative Mek1). EGF treatment caused a threefold increase in cell proliferation in mock and in control adenovirus-infected MCF10A cells (Fig. 3a). EGF-dependent cell growth was slightly reduced in MCF10A cells expressing either Erk2(AEF) (Fig. 3a) or Mek1(A) (data not shown). In contrast, MCF10A cells expressing Bmk1(AEF) failed to proliferate in response to EGF (Fig. 3a). These results demonstrate that the MAP kinase activity of Bmk1 is essential for EGF-mediated cell growth. A requirement for ERK activity has been demonstrated in triggering proliferative responses¹³. Our results do not contradict these findings as expression of the dominant negative forms of Erk2 or Mek1 are unlikely to completely block the ERK signalling pathway that consists of redundant kinases. In support of this, we find that expression of either Erk2(AEF) or Mek1(A) blocked only about 70% of the endogenous EGF-induced ERK stimulation in MCF10A cells, as determined in kinase and reporter gene assays⁷ (data not shown). Conversely, expression of Bmk1(AEF) blocked at least 97 per cent of endogenous EGF-mediated Bmk1 activation (data not shown).

As mitogens affect cell-cycle progression, leading to cell proliferation, we investigated the effect of Bmk1 on progression of the cell through the cell cycle. HeLa cells were infected with recombinant adenovirus expressing dominant-negative forms of Erk2, Mek1 or Bmk1, synchronized in the cell cycle by applying a double thymidine block, and analysed for cell-cycle progression by flow cytometry at various times after release. As expected, mock and control adenovirus-infected cells both progressed normally through the cell cycle (Fig. 4a). There was a small but measurable inhibition of progression through S phase for cells expressing either Erk2(AEF) or Mek1(A). In contrast, expression of Bmk1(AEF) blocked progression of cells through S phase. To determine a role for Bmk1 in cell-cycle progression induced by growth factors, MCF10A cells were starved in growth-factor-deficient medium for 48 h and the proportion of cells progressing to S phase was measured 15 h after addition

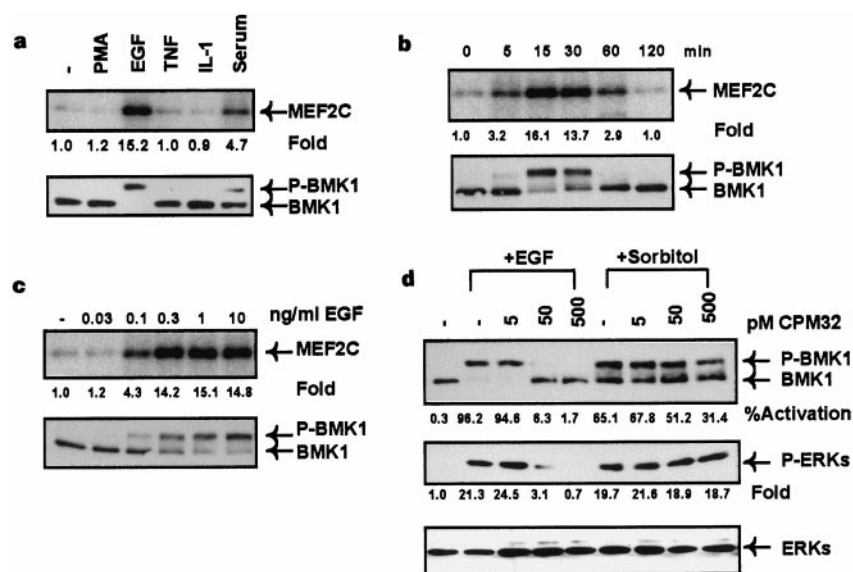


Figure 1 EGF is a potent inducer of Bmk1 activation. HeLa cells were serum-starved overnight and, after the treatment indicated, the activity of ERK and Bmk1 was determined. **a**, Cells were stimulated for 20 min with 100 nM PMA, 1 ng ml⁻¹ EGF, 100 ng ml⁻¹ TNF, 1 ng ml⁻¹ IL-1 or 10% serum. **b**, Cells were treated with 1 ng ml⁻¹

EGF for 0–120 min. **c**, Cells were treated with increasing doses of EGF for 20 min. **d**, Cells were treated with various amounts of CPM32 for 20 min and then exposed to 1 ng ml⁻¹ EGF or 0.4 M sorbitol for 20 min.

of growth factor. The addition of growth factor caused the fraction of cells in S phase to increase from about 5 to 40% for both mock and control adenovirus-infected cells (Fig. 4b). This growth-factor-dependent fraction of cells progressing to S phase was slightly reduced by the expression of Erk2(AEF). In contrast, expression of Bmk1(AEF) prevented cells from entering S phase following addition of growth factor. Taken together, our results show that Bmk1 activity is required for cells to enter the S phase of the cell cycle.

We have shown that EGF is a potent activator of Bmk1 and that Bmk1 activity is required for EGF-mediated cell proliferation and cell cycle progression. Mitogens mediate cellular responses by activating the expression of several immediate early response genes and we have previously shown that Bmk1 regulates the serum-induced expression of the immediate early response gene *c-jun*⁷. Thus, Bmk1 may mediate cellular responses to EGF by regulating the expression of essential early genes such as *c-jun*. It has been reported that microinjection of antibodies against the *c-jun* protein prevents serum-stimulated cells from entering the S phase¹⁴. As for all known EGF-mediated signals, we find that the tyrosine kinase activity of the EGF receptor is required for activation of Bmk1. However, although Ras is sufficient and necessary for EGF-mediated activation of both ERK and JNK MAP kinases^{10,15}, EGF-mediated activation of Bmk1 proceeds independently of Ras.

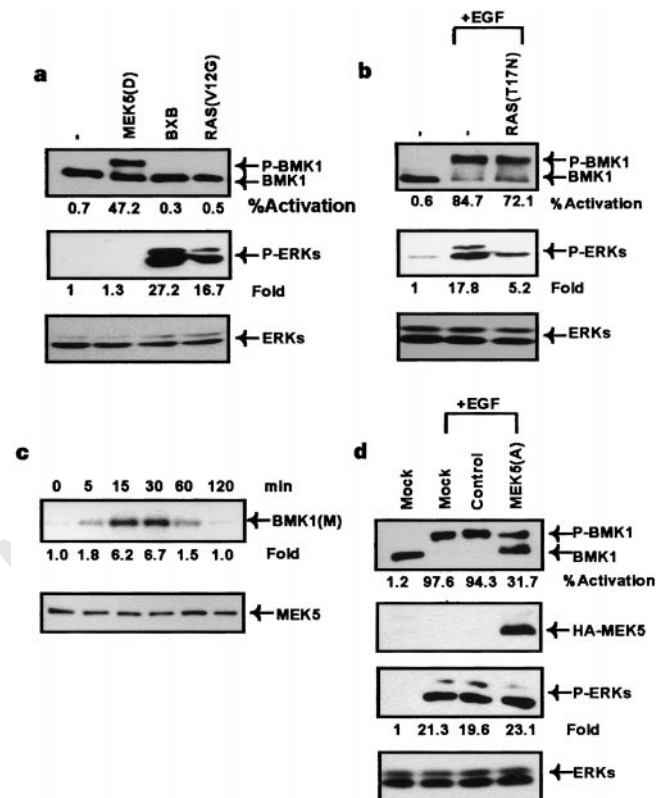


Figure 2 EGF activates Bmk1 through a Ras-independent and Mek5-dependent pathway. **a**, HeLa cells were transfected with expression vectors encoding Mek5(D), Raf-1(BXB) or Rad(V12G) and the activity of endogenous Bmk1 and ERK was determined 48 h later. **b**, HeLa cells were transfected with an expression vector encoding Ras(T17N) and treated 48 h later with 1 ng ml⁻¹ EGF for 20 min. **c**, Serum-starved HeLa cells were treated with 1 ng ml⁻¹ of EGF for 0–120 min and endogenous Mek5 activity was measured by an immune complex protein kinase assay using Bmk1(M) as a substrate. Western blot analysis of immunoprecipitated Mek5 is shown in the lower gel. **d**, HeLa cells were infected as indicated and treated 48 h post-infection with 1 ng ml⁻¹ EGF for 20 min. The activity of endogenous Bmk1 and ERK was measured as described in Methods. Second panel, western blot showing the expression of haemagglutinin (HA)-tagged Mek5(A) detected with anti-HA antibody 12CA5 (from I. Wilson).

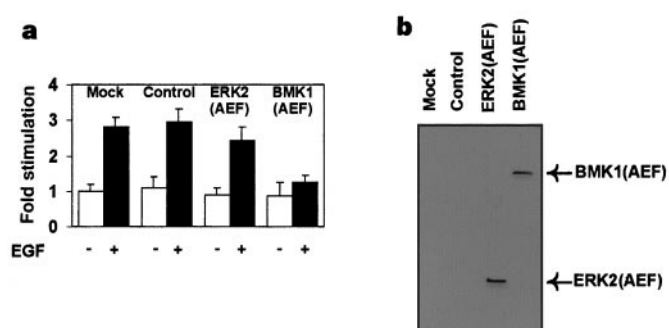


Figure 3 Bmk1 is required for EGF-dependent cell proliferation. **a**, MCF10A cells were infected as indicated and treated with EGF. Bars indicate relative growth stimulation compared with untreated cells, whose value was taken as unity. Error bars indicate standard deviations. **b**, Western blot showing the expression of Flag-tagged Bmk1(AEF) or Flag-tagged Erk2(AEF) detected by the anti-Flag antibody M2 (IBI-Kodak).

Our findings indicate that EGF-induced Bmk1 activation could be part of a signalling pathway distinct from that used by other MAP kinases such as ERK and JNK. In support of this, we have found that the MAP-kinase kinase Mek5 is a specific effector of EGF-mediated Bmk1 activation. Finally, as deregulated EGF-receptor activity is a hallmark of many types of tumour^{16–18}, Bmk1 may provide a drug target for the uncontrolled growth of malignant cells.

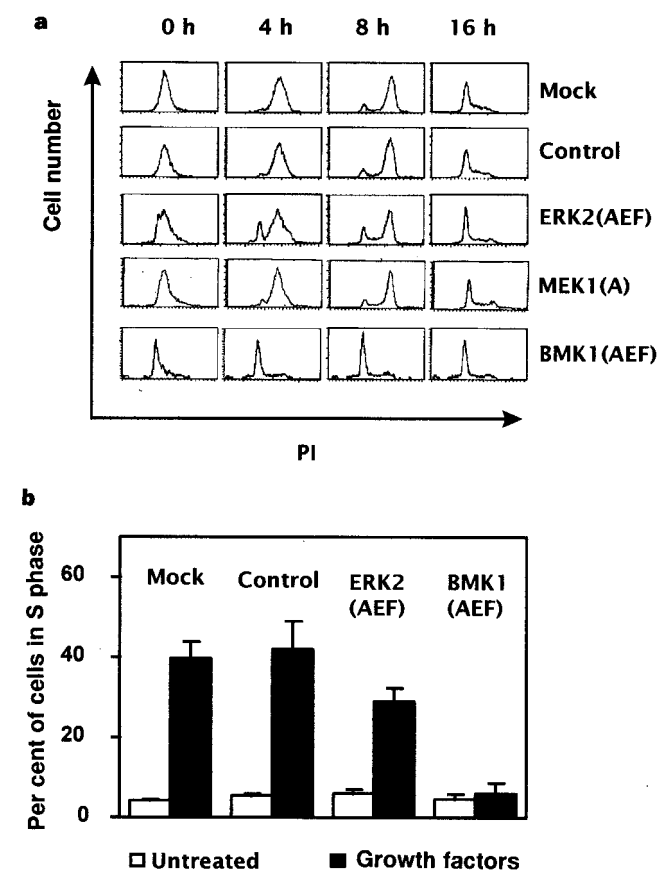


Figure 4 Bmk1 is required for entry of cells into S phase. **a**, HeLa cells were infected as indicated, synchronized at the G1-to-S phase boundary by a double thymidine block, stained with propidium iodide and analysed by flow cytometry at various times after release. **b**, MCF10A cells were infected, starved in growth-factor-deficient medium, and in some cases were treated with medium containing growth factors. The percentage of cells in S phase was measured in triplicate by flow cytometric analysis of total cellular DNA.

Methods

Immunoprecipitation, protein kinase assay and western blotting. Bmk1 and Mek5 were immunoprecipitated from solubilized⁷ HeLa cells as described¹⁹. Bmk1 kinase activity was assayed by an immune complex protein kinase assay and quantified as described⁷. Phosphorylation of Bmk1 results in a shift in its electrophoretic mobility detectable by western blotting, and we have consistently seen a correlation between the shifted fraction of phosphorylated Bmk1 and the activity of Bmk1 as measured by protein kinase assay^{6,7} (Fig. 1a–c). On western blots, the relative percentage activation of Bmk1 was calculated by dividing the density of the upper phosphorylated Bmk1 band by the total density of the upper and lower Bmk1 bands, and multiplying by 100. Activation of endogenous ERK was measured in cell lysates by western blotting using an anti-ERK antibody (Santa Cruz Biotechnology) or a phospho-specific anti-ERK antibody (New England Biolabs), respectively. Relative activation was determined by measuring the density of each phosphorylated ERK band relative to that of untreated cells whose activity was taken as unity.

Expression vectors and recombinant adenoviruses. Raf-1(BXB), Ras(V12G) and Ras(T17N) pSRα vectors were gifts from M. Karin^{10,20,21}. The cDNAs for the various forms of Erk2, Mek1, Mek5 and Bmk1 (ref. 7), were subcloned into the *HindIII/XbaI* site of the vector pAd/RSV²². Recombinant adenoviruses expressing the various forms of Erk2, Mek1, Mek5 and Bmk1 were generated as described^{23–25}. The control adenovirus was generated using the empty vector. Viral DNA was isolated and insertion of the appropriate cDNA was confirmed by PCR²². High-titre recombinant adenoviral stocks of ~10¹¹ PFU ml⁻¹ were produced in 293 cells and purified by density gradient ultracentrifugation²³.

Proliferation and cell cycle analysis. MCF10A cells were grown in DMEM/F12 medium supplemented with 10% FCS, 10 ng ml⁻¹ EGF, 5 μg ml⁻¹ insulin and 0.5 μg ml⁻¹ hydrocortisone. To ensure optimal expression of recombinant proteins, MCF10A cells were infected with recombinant adenovirus at an MOI of 10³ PFU per cell. Two days after infection, cells were plated into 96-well dishes at 2 × 10³ cells per well using serum-free medium (DMEM/F12) containing 1 mg fetuin, 10 μg transferrin, 5 μg insulin and 0.5 μg hydrocortisone per ml with or without 100 ng ml⁻¹ of EGF. Growth was monitored 3 days later using an MTT-based cell-proliferation assay (Boehringer) as specified by the supplier. Preliminary experiments established a linear correlation between the absorbance measured in the MTT assay and actual cell number over a range of 200–50,000 MCF10A cells per well. All assays were done in triplicate within this linear range.

HeLa cells were incubated with or without recombinant adenovirus at an MOI of 10³ PFU per cell. 48 h after infection, cells were synchronized at the G1–S phase boundary by a double thymidine block as described²⁶. Cells were released from this double thymidine block for various times and analysed by flow cytometry after staining nuclear DNA with propidium iodide²⁷. In a separate experiment, MCF10A cells were starved for 48 h in unsupplemented DMEM/F12 medium. Some cells were exposed to growth factors by addition of DMEM/F12 medium containing 10% fetal calf serum, 10 ng ml⁻¹ EGF, 5 μg ml⁻¹ insulin and 0.5 μg ml⁻¹ hydrocortisone. After 15 h, the percentage of cells in S phase was measured in triplicate by adding 10 μM bromodeoxyuridine (Boehringer) directly to the culture medium for 45 min. The cells were then collected, stained with propidium iodide and anti-BrdU fluorescein isothiocyanate (FITC)-conjugated antibody (Boehringer). The fraction of cells in S phase was determined by flow cytometry^{28,29}.

Received 26 March; accepted 7 August 1998.

- Yarden, Y. & Ullrich, A. Growth factor receptor tyrosine kinases. *Annu. Rev. Biochem.* **57**, 443–478 (1988).
- Ullrich, A. & Schlessinger, J. Signal transduction by receptors with tyrosine kinase activity. *Cell* **61**, 203–212 (1990).
- Su, B. & Karin, M. Mitogen-activated protein kinase cascades and regulation of gene expression. *Curr. Opin. Immunol.* **8**, 402–411 (1996).
- Johnson, G. L. & Vaillancourt, R. R. Sequential protein kinase reactions controlling cell growth and differentiation. *Curr. Opin. Cell Biol.* **6**, 230–238 (1994).
- Katz, M. E. & McCormick, F. Signal transduction from multiple Ras effectors. *Curr. Opin. Genet. Dev.* **7**, 75–79 (1997).
- Abe, J.-I., Kushuhara, M., Ulevitch, R. J., Berk, B. C. & Lee, J.-D. Big mitogen-activated protein kinase 1 (BMK1) is a redox-sensitive kinase. *J. Biol. Chem.* **271**, 16586–16590 (1996).
- Kato, Y. *et al.* BMK1/ERK5 regulates serum-induced early gene expression through transcription factor MEF2C. *EMBO J.* **16**, 7054–7066 (1997).
- Li, J. J. & Herskowitz, I. Isolation of ORC6 a component of the yeast origin recognition complex by a one-hybrid system. *System* **262**, 1870–1874 (1993).

- Lange-Carter, C. A. & Johnson, G. L. Ras-dependent growth factor regulation of MEK kinase in PC12 cells. *Science* **265**, 1458–1461 (1994).
- Minden, A. *et al.* Differential activation of ERK and JNK mitogen-activated protein kinases by Raf-1 and MEKK. *Science* **266**, 1719–1723 (1994).
- Zhou, G., Bao, Z. Q. & Dixon, J. E. Components of a new human protein kinase signal transduction pathway. *J. Biol. Chem.* **270**, 12665–12669 (1995).
- Soule, H. D. *et al.* Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10. *Cancer Res.* **50**, 6075–6086 (1990).
- Pages, G. *et al.* Mitogen-activated protein kinases p42mapk and p44mapk are required for fibroblast proliferation. *Proc. Natl Acad. Sci. USA* **90**, 8319–8323 (1993).
- Kovary, K. & Bravo, R. The jun and fos protein families are both required for cell cycle progression in fibroblasts. *Mol. Cell Biol.* **11**, 4466–4472 (1991).
- Barr, M. M., Tu, H., Van Acist, L. & Wigler, M. Identification of Ste4 as a potential regulator of Byr2 in the sexual response pathway of *Schizosaccharomyces pombe*. *Mol. Cell Biol.* **16**, 5597–5603 (1996).
- Libermann, T. A. *et al.* Amplification, enhanced expression and possible rearrangement of EGF receptor gene in primary human brain tumours of glial origin. *Nature* **313**, 144–147 (1985).
- Velu, T. J. *et al.* Epidermal-growth-factor-dependent transformation by a human EGF receptor proto-oncogene. *Science* **238**, 1408–1410 (1987).
- Merlino, G. T. *et al.* Amplification and enhanced expression of the epidermal growth factor receptor gene in A431 human carcinoma cells. *Science* **224**, 417–419 (1984).
- Han, J. *et al.* Characterization of the structure and function of a novel MAP kinase kinase (MKK6). *J. Biol. Chem.* **271**, 2886–2891 (1996).
- White, M. A. *et al.* Multiple Ras functions can contribute to mammalian cell transformation. *Cell* **80**, 533–541 (1995).
- Minden, A., Lin, A., Claret, F.-X., Abo, A. & Karin, M. Selective activation of the JNK signaling cascade and c-Jun transcriptional activity by the small GTPases Rac and Cdc42Hs. *Cell* **81**, 1147–1157 (1995).
- Huang, S. *et al.* Apoptosis signaling pathway in T cells is composed of ICE/Ced-3 family proteases and MAP kinase kinase 6b. *Immunity* **6**, 739–749 (1997).
- Chang, M. W., Barr, E., Lu, M. M., Barton, K. & Leiden, J. M. Adenovirus-mediated over-expression of the cyclin/cyclin-dependent kinase inhibitor p21 inhibits vascular smooth muscle cell proliferation and neointima formation in the rat carotid artery model of balloon angioplasty. *J. Clin. Invest.* **96**, 2260–2268 (1995).
- Bett, A. J., 'Iac'ara, J., Prevec, L. & Graham, F. L. An efficient and flexible system for construction of adenovirus vectors with insertions or deletions in early regions 1 and 3. *Proc. Natl Acad. Sci. USA* **91**, 8802–8806 (1994).
- Jomary, C. *et al.* Adenovirus-mediated gene transfer to murine retinal cells *in vitro* and *in vivo*. *FEBS Lett.* **347**, 117–122 (1994).
- Stein, G. S. *et al.* in *A Laboratory Handbook* (ed. Celis, J.) 282–287 (Danish Centre for Human Genome Research, Academic, San Diego, 1994).
- Draetta, G. & Beach, D. Activation of cdc2 protein kinase during mitosis in human cells: cell cycle-dependent phosphorylation and subunit rearrangement. *Cell* **54**, 17–26 (1988).
- Hoy, C. A., Seamer, L. C. & Shimke, R. T. Thermal denaturation of DNA for immunochemical staining of incorporated bromodeoxyuridine (BrdUrd): critical factors that affect the amount of fluorescence and the shape of BrdUrd/DNA histogram. *Cytometry* **10**, 718–725 (1989).
- Kastan, M. B., Onyekwere, O., Sidransky, D., Vogelstein, B. & Craig, R. W. Participation of p53 protein in the cellular response to DNA damage. *Cancer Res.* **51**, 6304–6311 (1991).

Acknowledgements. We thank M. Karin for Ras(V12G) and Ras(T17N) expression vectors, and B. Chastain for secretarial assistance. This work was supported by grants from the NIH (to J.-D.L. and R.J.U.) and the American Heart Association (to J.-D.L.) and by postdoctoral fellowships from the American Heart Association, California Affiliate (R.I.T.) and the Research Center for Infectious Disease, Aichi Medical University (Y.K.).

Correspondence and requests for materials should be addressed to J.-D.L. (e-mail: jdlee@scripps.edu)

Suppression of auxin signal transduction by a MAPK cascade in higher plants

Yelena Kovtun, Wan-Ling Chiu, Weike Zeng & Jen Sheen

Department of Genetics, Harvard Medical School, and Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA

The plant hormone auxin activates many early response genes that are thought to be responsible for diverse aspects of plant growth and development¹. It has been proposed that auxin signal transduction is mediated by a conserved signalling cascade consisting of three protein kinases: the mitogen-activated protein kinase (MAPK), MAPK kinase (MAPKK) and MAPKK kinase (MAPKKK)². Here we show that a specific plant MAPKKK, NPK1 (ref. 3), activates a MAPK cascade that leads to the suppression of early auxin response gene transcription. A mutation in the kinase domain abolishes NPK1 activity, and the presence of the carboxy-terminal domain diminishes the kinase activity. Moreover, the effects of NPK1 on the activation of a MAPK and the repression of early auxin response gene transcription are specifically eliminated by a MAPK phosphatase⁴. Transgenic tobacco plants overexpressing the NPK1 kinase domain produced seeds defective in embryo and endosperm development.

These results suggest that auxin sensitivity may be balanced by antagonistic signalling pathways that use a distinct MAPK cascade in higher plants.

A transient expression system using maize mesophyll protoplasts has been developed to elucidate the molecular mechanisms of gene expression and intracellular signal transduction stimulated by sugars, light and the plant hormone abscisic acid⁵⁻⁹. To determine whether the system is suitable for the study of auxin signalling, we tested the auxin inducibility of a well-characterized early response gene promoter, GH3 (ref. 10), in maize protoplasts. The protoplasts transfected with a construct carrying the coding region of a green fluorescent protein (sGFP)¹¹ driven by the soybean GH3 promoter (GH3-sGFP) showed bright fluorescence on induction with different active auxin forms, NAA (Fig. 1a) or IAA (data not shown). In contrast, auxin did not affect the expression of sGFP controlled by the maize chlorophyll *a/b* binding protein gene promoter (CAB5)⁵ (Fig. 1a). The auxin inducibility of the GH3 promoter in maize protoplasts was verified by using another reporter gene encoding *Escherichia coli* β -glucuronidase (GUS) (Fig. 1b).

To confirm that the early auxin responses are conserved among higher plants, we tested an auxin responsive element, ER7 (ref. 12), which was found in most early auxin response gene promoters^{1,12}. The ER7 promoter element was inserted upstream of the GUS gene driven by a 35S minimal promoter. The ER7 synthetic promoter conferred auxin inducibility in the system, whereas the mutated ER7 element or the minimal promoter alone could not be induced by auxin (Fig. 1b). The data demonstrate that maize mesophyll

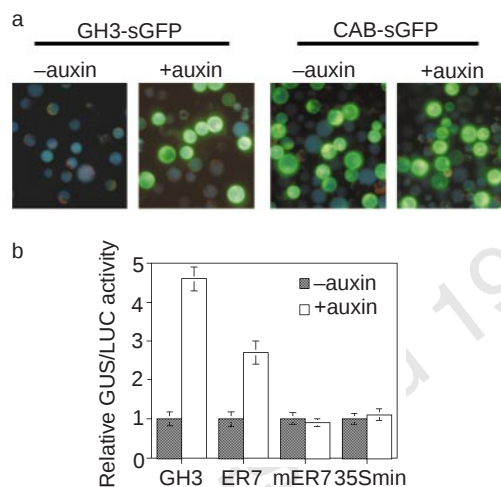


Figure 1 Auxin responses in maize protoplasts. **a**, Visualization of auxin responses. Protoplasts were transfected with plasmid DNA carrying GH3-sGFP or CAB5-sGFP and incubated without or with auxin. **b**, The GH3 promoter and the ER7 auxin responsive element are regulated in maize protoplasts. The protoplasts were transfected with plasmid DNA carrying GH3-GUS (GH3), ER7-GUS (ER7), mutated ER7-GUS (mER7), or GUS controlled by the CaMV 35S minimal (-72) promoter (35Smin). The protoplasts were incubated without or with auxin. GUS/LUC activity of the protoplasts transfected with each construct and incubated without auxin was set to 1.

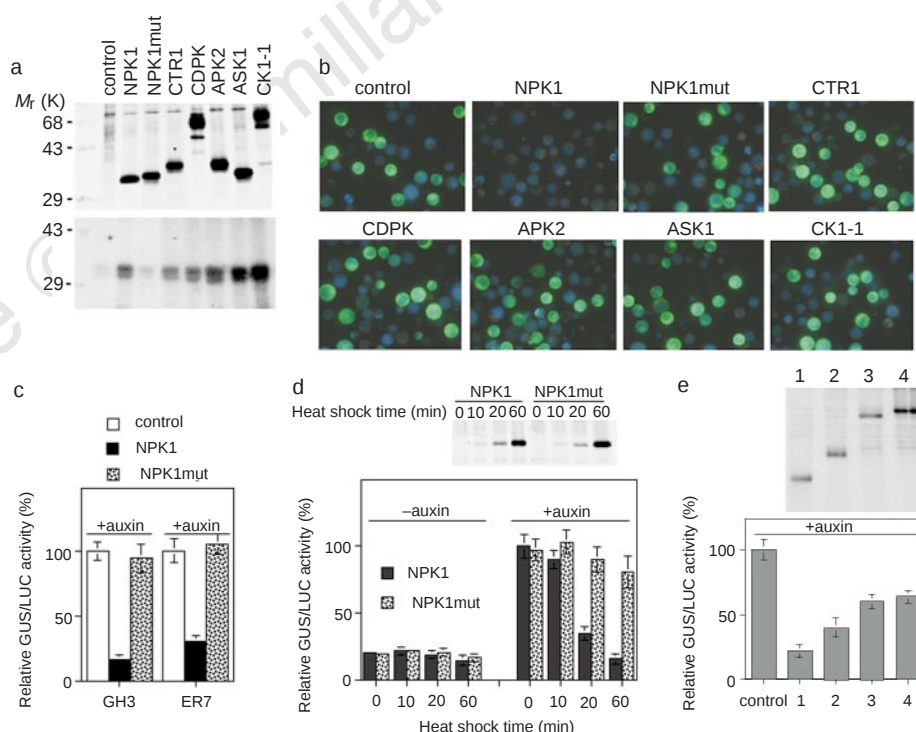


Figure 2 NPK1 represses the auxin-inducible promoters. **a**, Analyses of protein kinases in maize protoplasts. Protoplasts were transfected with vector DNA (control) or an effector construct carrying various protein kinases (NPK1, CTR1, CDPK, APK2, ASK1, CK1-1) or NPK1 K109M mutant (NPK1mut). Top, the effector protein expression level; bottom, effector kinase activity. **b**, NPK1 represses the GH3 promoter. Protoplasts were co-transfected with the GH3-sGFP reporter and an effector construct carrying the kinases indicated or vector DNA (control), and incubated with auxin to induce the GH3 promoter. **c**, NPK1 represses auxin-responsive promoters. Protoplasts were co-transfected with GH3-GUS (GH3), or ER7-GUS (ER7) reporter and an effector construct carrying NPK1 or NPK1mut, or vector DNA (control). **d**, NPK1 protein levels correlate with the inhibitory effect. The expression of wild-type (NPK1) or mutated (NPK1mut) kinase domain is

controlled by a heat-shock-inducible promoter (HSP)¹⁴. Protoplasts were co-transfected with the GH3-GUS reporter and HSP-NPK1 or HSP-NPK1mut effector. Expression of the NPK1 or NPK1mut protein was induced at 42°C. The protoplasts from each treatment were divided to determine the effector protein expression level (top) and to measure GUS/LUC activity (bottom). **e**, Analysis of the putative regulatory domains of NPK1. Protoplasts were co-transfected with the GH3-GUS reporter and vector DNA (control), or one of the effector constructs carrying the coding region of NPK1 kinase domain only (lane 1), N terminus and kinase domain (lane 2), kinase domain and C terminus (lane 3) or full-length NPK1 protein (lane 4). The protoplasts were divided to determine the effector protein expression level (top) and to measure GUS/LUC activity (bottom).

protoplasts respond to physiological levels of auxin, and that the early auxin responses are probably conserved in both monocotyledonous and dicotyledonous plants.

It has been reported that auxin activates a MAPK in tobacco². The presence of NPK1, a tobacco MAPKKK, in fast-growing cells³ suggested that NPK1 could mediate auxin signalling. To determine whether NPK1 is involved in auxin signal transduction, we tested the effect of NPK1 on the GH3 promoter and the ER7 promoter element. The NPK1 kinase domain was tagged with a double haemagglutinin epitope (DHA)⁸ and cloned into a plant expression vector with a derivative of the CaMV35S promoter (not affected by auxin)^{6,9}. The NPK1 construct was co-transfected with the GH3-sGFP or GH3-GUS construct into maize protoplasts. Expression and kinase activity of the NPK1 protein were confirmed in transfected protoplasts (Fig. 2a). Surprisingly, NPK1 blocked auxin activation of the GH3 promoter and the ER7 promoter element (Fig. 2b, c). However, the activities of many auxin-insensitive promoters, including the promoters of CAB, actin, ubiquitin, and CaMV35S genes, were not affected by NPK1 (data not shown). To show that the kinase activity of NPK1 is necessary for the repression, a kinase-inactive mutant (K109M) was created by site-directed mutagenesis. The mutation did not affect the expression of the NPK1 protein but abolished the protein kinase activity (Fig. 2a) and the negative effect of NPK1 on the GH3 promoter and the ER7 promoter element in the presence of auxin (Fig. 2b, c).

To demonstrate that the inhibitory effect was specific to NPK1, we tested the effect of other serine/threonine protein kinases that belong to five different classes, including another plant MAPKKK, *Arabidopsis* CTR1 (ref. 13). These five kinases were expressed and displayed protein kinase activity in maize protoplasts (Fig. 2a), but did not block auxin signalling (Fig. 2b).

To test the effect of different NPK1 protein levels on the GH3 promoter activity, we used a heat-shock promoter¹⁴ to control the amount of the NPK1 protein in the transfected protoplasts. The active and mutated NPK1 protein levels correlated well with the duration of heat shock (Fig. 2d). In the absence of auxin, NPK1 could not activate the GH3 promoter. In the presence of auxin, the inverse correlation between the NPK1 protein level and GH3 promoter activity supports the idea that NPK1 acts as a negative regulator in auxin signal transduction (Fig. 2d).

NPK1 consists of a conserved kinase domain, a short amino terminus and a long carboxy-terminal sequence³. To investigate the function of the regions outside the kinase domain, we created several NPK1 deletions and tested their effect on GH3 promoter activity. All constructs showed similar levels of protein expression in transfected maize protoplasts (Fig. 2e, top). The kinase region alone or the kinase domain with the N terminus inhibited the GH3 promoter more effectively than the N-terminal deletion or the full-length NPK1 (Fig. 2e, bottom). The data are consistent with the previous observation that the NPK1 protein lacking the C terminus was more effective in the yeast complementation assay³. It is likely that the regions outside the NPK1 kinase domain regulate kinase activity, and signals antagonizing auxin responses are required to completely activate the full-length NPK1. Because several signals, such as cell-wall-derived oligosaccharides, cytokinin, ABA and ethylene¹⁵, can interfere with auxin responses in plants, future work will be directed towards identifying the physiologically relevant signals that control NPK1 activity and counteract the auxin responses.

Because NPK1 is a MAPKKK, it is expected to induce a protein phosphorylation cascade resulting in the activation of a MAPK. Although several plant MAPKs have been shown to be induced by stress, hormone and elicitor signals^{2,16,17}, their activation by a phosphorylation cascade has not been demonstrated in plant cells. We performed an in-gel MAPK activity assay^{2,17} with extracts prepared from transfected protoplasts (Fig. 3a, top). Protoplasts transfected with the NPK1 kinase domain had significantly higher

endogenous MAPK activity (relative molecular mass, M_r , of 44K) than protoplasts transfected with the mutated NPK1 or an empty vector. The expression of the full-length NPK1 increased the putative MAPK activity only marginally, consistent with the observation that the full-length protein repressed the GH3 promoter mildly (Fig. 2e). CTR1 also activated an endogenous kinase (Fig. 3a), implying the existence of another unrelated MAPK cascade in maize protoplasts.

To verify that NPK1 expression resulted in the activation of a MAPK, we performed kinase activity assays with the proteins immunoprecipitated with an antibody raised against two conserved domains of a mammalian MAPK. The MAPK activity of the protoplasts transfected with the active NPK1 was significantly higher than that of the cells transfected with the mutated NPK1 (Fig. 3a, bottom). These data are consistent with the results of the in-gel MAPK activity assay (Fig. 3a, top), and indicate that tobacco NPK1 can induce a kinase cascade and activate an endogenous MAPK in maize protoplasts.

To determine whether the 44K MAPK is involved in the repression of early auxin response genes, we tested the effect of a MAPK-specific phosphatase (MKP). Protein phosphatases that can specifically dephosphorylate and inactivate MAPKs have been reported in a variety of eukaryotes and are evolutionarily conserved⁴. A mouse MKP1 (ref. 4), highly specific to MAPKs, was cloned into the plant expression vector and expressed in maize protoplasts. The expression of MKP1, but not other plant protein phosphatases (PP) that belong to the three serine/threonine classes (PP1, PP2A⁶, and PP2C⁹; Fig. 3b, top), resulted in the complete elimination of the NPK1 effects, including the NPK1-dependent activation of a MAPK (Fig. 3b, bottom) and the repression of the GH3 promoter (Fig. 3c).

To assess the function of NPK1 at a whole-plant level, we generated transgenic tobacco plants overexpressing the NPK1

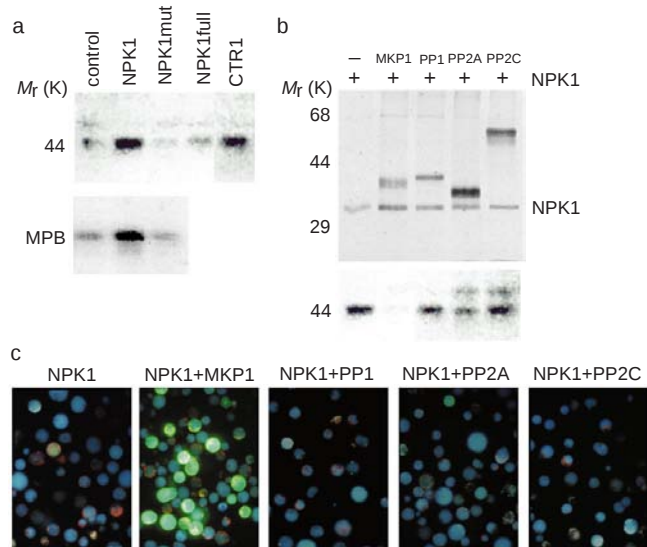


Figure 3 NPK1 activates a MAPK. **a**, MAPK activity assays. Maize protoplasts were transfected with vector DNA (control) or constructs carrying the coding region of NPK1 kinase domain (NPK1), mutated NPK1 kinase domain (NPK1mut), full-length NPK1 (NPK1full) or CTR1 kinase domain (CTR1). In-gel (top) and immunocomplex (bottom) kinase activity assays were performed with cell lysates from the transfected protoplasts. **b**, MAPK phosphatase inactivates NPK1-induced MAPK. Protoplasts were co-transfected with NPK1 and MAPK phosphatase (MKP1) or serine/threonine protein phosphatase (PP1, PP2A, PP2C) effector constructs. The transfected protoplasts were divided to determine the effector protein expression level (top) and to perform the in-gel kinase activity assay (bottom). **c**, MKP1 abolishes NPK1 repression of auxin-inducible transcription. Protoplasts were co-transfected with the GH3-sGFP reporter and NPK1, or NPK1 plus the phosphatases indicated, and incubated in a medium with auxin.

kinase domain. It was anticipated that overexpression of NPK1, as an auxin antagonist, could be lethal. Indeed, in the strongest transgenic line about 75% of the seeds never germinated, whereas seeds from the wild-type control (Fig. 4a) and many other tobacco lines carrying other transgenes (data not shown) germinated normally. Closer examination showed that some transgenic seeds had visible defects in embryo and endosperm (Fig. 4b); for example, many transgenic seeds were hollow with little endosperm and retarded embryos, and in the most severe cases transgenic seeds completely lacked both endosperm and embryo (Fig. 4b). The number of the defective seeds in each line (Fig. 4a) correlated with the level of the transgene expression (Fig. 4c), suggesting that the seed phenotype was due to transgene expression and enhanced NPK1-dependent MAPK activity. Similarly, ectopic activation of a MAPK cascade can disrupt embryo development by inducing mitotic arrest at the M phase in *Xenopus*¹⁸.

Surprisingly, the surviving NPK1 transgenic seeds produced

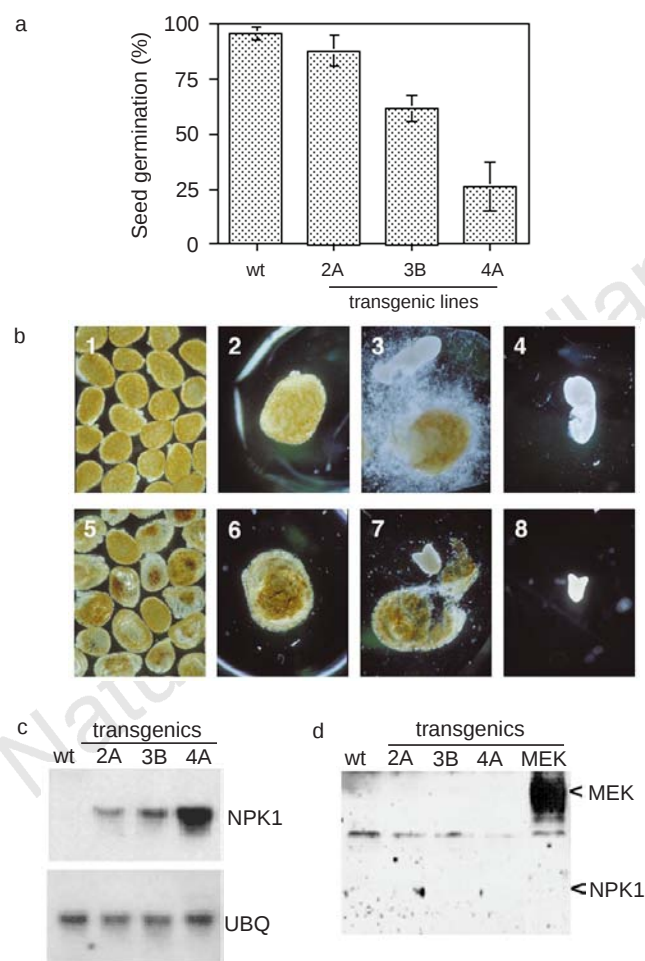


Figure 4 Analyses of NPK1 transgenic tobacco plants. **a**, Seed germination. Wild-type (wt) and three independent transgenic lines (2A, 3B and 4A) were examined. The results are the means of triplicate samples, 100 seeds each, $\pm$ s.d. **b** NPK1 transgenic seeds display defects in embryo and endosperm. The wt (1–4) and 4A (5–8) seeds were soaked for 24 hours in water. The seeds are shown as a population (1 and 5), a typical single seed (2 and 6) and dissected (3 and 7). A representative embryo from wt (4) or 4A (8) is shown. At least 10 seeds from each population were analysed. **c**, RNA blot analysis of the transgene expression. The NPK1 probe hybridized with the transgene RNA only. Ubiquitin (UBQ)³⁰ expression was used as a control. **d**, Protein blot analysis of the transgene expression. The same amount of protein (50 μ g per lane) was fractionated by SDS-PAGE (12%). An anti-haemagglutinin antibody was used to detect DHA-tagged transgene proteins. A tobacco transgenic line overexpressing a DHA-tagged MEK protein (MEK) was used as a positive control.

plants of wild-type appearance. These plants had a high level of ectopic NPK1 mRNA (Fig. 4c) but we could not detect the ectopic DHA-tagged NPK1 protein after several protein blot analyses, whereas the control transgenic line expressing the DHA-tagged MEK1 showed a strong signal (Fig. 4d). We hypothesize that the truncated NPK1 protein is unstable and cannot accumulate to a level sufficient to cause grossly abnormal phenotypes. This is in agreement with a recent report that MEKK1, a mammalian MAPKKK, is degraded rapidly after processing and activation¹⁹. In tobacco cells the NPK1 protein is subjected to a fast turnover after activation at the end of M phase in the cell cycle²⁰, and is only detectable at low levels in dividing cells²¹. Thus, accumulation of NPK1 protein might be tightly regulated. This could explain why we observed the most dramatic effect of NPK1 during embryogenesis and seed development, when rapid cell divisions occur and more NPK1 proteins could accumulate to block cell-cycle progression. Recent analyses of auxin-resistant mutants in *Arabidopsis* suggested a crucial role for protein degradation in auxin signalling and cell-cycle control. For example, several auxin-resistant mutants (*axr1*, *tir1*) show defects in protein degradation processes^{22,23}. Many auxin-inducible proteins, such as SAUR and Aux/IAA, are highly unstable, and some of them function as negative regulators of auxin-mediated transcription^{1,24,25}. Our data provide another indication that cell cycle, protein turnover and auxin signalling are connected.

It has been shown that conserved MAPK cascades mediate several vital functions in mammals and yeast, such as cell proliferation, cell death and stress responses²⁶. Although many plant MAPK, MAPKK and MAPKKK homologues have been identified on the basis of sequence conservation and functional complementation in yeast, their precise physiological functions in plants are mostly unknown¹⁶. Here we demonstrate that a plant MAPKKK can activate a MAPK cascade controlling gene expression, act as a negative regulator in the auxin signal transduction pathway, and potentially disturb embryogenesis in transgenic tobacco. The finding of three NPK1-like protein kinases in *Arabidopsis*²⁷ suggests that this distinct MAPKKK is conserved in higher plants and that a selection for the desired NPK1 mutants could be problematic owing to potential functional redundancy or lethality. Our data suggest that the tightly regulated protein turnover might complicate the use of transgenic plants in the functional analysis of the NPK1-like MAPKKKs. The protoplast transient expression system offers an alternative strategy for investigating the functions of known genes, especially those encoding tightly regulated proteins, with great specificity and flexibility.

Methods

Reporter constructs. The 749-base pair (bp) soybean GH3 promoter¹⁰ was fused to a synthetic green fluorescent protein gene¹¹ to create the GH3–sGFP reporter construct. A synthetic ER7 element, TTGTCTCCCAAAGGGAGACA, or mutated ER7, TTGTCTCCCAAAGGGAGATAA¹², was inserted in front of the CaMV 35S minimal promoter (–72)⁶. The synthetic promoters were fused to a GUS gene to create ER7–GUS and mER7–GUS reporter constructs. Three clones of each construct gave identical results when tested for auxin induction.

Effector constructs. NPK1, NPK1 deletions and CTR1 were obtained by the polymerase chain reaction (PCR) from tobacco cDNA and an *Arabidopsis* cDNA library, respectively. The kinase-inactive NPK1 mutant (K109M) was generated by PCR using the following primers: TACTCGTATAATG–GAGGTTTCGAT and CGCAATCGAAACCTCCaTTATAGCGAGTA. The mutation was confirmed by DNA sequencing. All PCR products, and the coding regions of protein kinases from *Arabidopsis* (CDPK (ref. 8), APK2 (ref. 8), ASK1 (ref. 8) and CK1-1 (ref. 28)) and protein phosphatases (mouse MPK1 (ref. 4), maize PP1 (ref. 6), maize PP2A (ref. 6) and *Arabidopsis* PP2C (ref. 9)) were tagged with two copies of the haemagglutinin epitope (DHA) and inserted into a plant expression vector containing the 35S4PPDK promoter and the *nos* terminator^{8,9}. Three to four independent effector clones were tested and gave identical results.

Protoplast transient expression. The preparation and electroporation of etiolated maize mesophyll protoplasts have been described^{5,6}. In each electroporation, 2×10^5 protoplasts were transfected with 30 μ g of plasmid DNA carrying a reporter construct alone or with 30 μ g of plasmid DNA carrying an effector construct or a vector DNA control. The transfected protoplasts were incubated in medium (5×10^4 ml⁻¹) with (+ auxin) or without (- auxin) 1 μ M NAA for 14 hours in the dark. GFP fluorescence was observed under ultraviolet irradiation by fluorescent microscopy¹⁴. A construct carrying the maize CAB5 promoter⁵ fused to the luciferase gene (CAB-LUC) was used as an internal control in each transfection where the GH3-GUS or ER7-GUS reporter construct was tested. In each sample the GUS activity⁵ of the cell lysate was divided by the LUC activity⁸, thereby normalizing the data for variation in experimental conditions (number of cells, transformation efficiency and cell viability). Our results are shown as the means of triplicate samples $\pm$ the standard deviation. All transient expression experiments were repeated three times with similar results.

Effector protein expression level. Transfected maize protoplasts (10^5) were incubated for 5 hours with [³⁵S]-methionine (200 μ Ci ml⁻¹). The DHA-tagged effectors were immunoprecipitated with an anti-haemagglutinin antibody⁸, separated by SDS-PAGE (10%) and visualized by fluorography.

Protein kinase activity assays. Cell lysates from 10^5 transfected protoplasts were used for immunoprecipitation with an anti-haemagglutinin antibody⁸. The immunoprecipitated proteins were assayed with [γ -³²P]ATP and casein as a substrate. The [³²P]-casein was separated by SDS-PAGE (10%) and visualized by autoradiography.

MAPK activity assays. The transfected protoplasts (10^5) were incubated for 5 h before collection. For the in-gel kinase activity assay, protoplasts extracts containing 7 μ g of protein were fractionated in 10% SDS-PAGE gel embedded with 0.25 mg ml⁻¹ of myelin basic protein (MBP), a MAPK substrate. The protein denaturing, renaturing and kinase activity in the gel were performed as described¹⁷. For immunocomplex kinase activity assay, maize endogenous MAPKs were immunoprecipitated from protoplast lysates with an anti-ERK (pan-ERK) antibody (Transduction Laboratory)⁸ and assayed for MAPK activity with MBP as a substrate¹⁷. The [³²P]-MBP was separated by SDS-PAGE (15%) and visualized by autoradiography.

Transgenic plants. A construct including the 35S4PPDK promoter^{6,8,9}, the kinase domain of NPK1, a DHA tag and a *nos* terminator was inserted into the pART27 binary vector²⁹. The resulting plasmid was introduced into *Agrobacterium tumefaciens* EHA105 and the transformation was performed with *Nicotiana tabacum* SR1 leaf discs¹¹. Several kanamycin-resistant plants were selected from three independent transformation experiments. The seeds were examined by light microscopy. Seedlings two weeks old were used for RNA blot and protein blot analyses³⁰.

Received 23 June; accepted 10 August 1998.

1. Abel, S. & Theologis, A. Early genes and auxin action. *Plant Physiol.* **111**, 9–17 (1996).
2. Mizoguchi, T. *et al.* Characterization of two cDNAs that encode MAP kinase homologues in *Arabidopsis thaliana* and analysis of the possible role of auxin in activating such kinase activities in cultured cells. *Plant J.* **5**, 111–122 (1994).
3. Banno, H. *et al.* NPK1, a tobacco gene that encoded a protein with a domain homologous to yeast BCK1, STE11, and Byr2 protein kinases. *Mol. Cell. Biol.* **13**, 4745–4752 (1993).
4. Sun, H. *et al.* MKP-1 (3CH134), an immediate early gene product, is a dual specificity phosphatase that dephosphorylates MAP kinase *in vivo*. *Cell* **75**, 487–493 (1993).
5. Sheen, J. Metabolic repression of transcription in higher plants. *Plant Cell* **2**, 1027–1038 (1990).
6. Sheen, J. Protein phosphatase activity is required for light-inducible gene expression in maize. *EMBO J.* **12**, 3497–3505 (1993).
7. Jang, J.-C. & Sheen, J. Sugar sensing in higher plants. *Plant Cell* **6**, 1665–1679 (1994).
8. Sheen, J. Ca²⁺-dependent protein kinases and stress signal transduction in plants. *Science* **274**, 1900–1902 (1996).
9. Sheen, J. Mutational analysis of protein phosphatase 2C involved in abscisic acid signal transduction in higher plants. *Proc. Natl Acad. Sci. USA* **95**, 975–980 (1998).
10. Hagen, G. *et al.* Auxin-induced expression of the soybean GH3 promoter in transgenic tobacco plants. *Plant Mol. Biol.* **17**, 567–579 (1991).
11. Chiu, W. *et al.* Engineered GFP as a vital reporter in plants. *Curr. Biol.* **6**, 325–330 (1996).
12. Ulmasov, T. *et al.* ARF1, a transcription factor that binds to auxin response elements. *Science* **276**, 1865–1868 (1997).
13. Kieber, J. J. *et al.* CTR1, a negative regulator of the ethylene response pathway in arabidopsis, encoded a member of the RAF family of protein kinases. *Cell* **72**, 427–441 (1993).
14. Sheen, J. *et al.* Green-fluorescent protein as a new vital marker in plant cells. *Plant J.* **8**, 777–784 (1995).
15. John, M. *et al.* Cell signalling by oligosaccharides. *Trends Plant Sci.* **2**, 111–115 (1997).
16. Hirt, M. Multiple roles of MAP kinases in plant signal transduction. *Trends Biol. Sci.* **2**, 11–15 (1997).
17. Zhang, S. & Klessig, D. Salicylic acid activates a 48-kD MAP kinase in tobacco. *Plant Cell* **9**, 809–824 (1997).
18. Takenaka, K. *et al.* Activation of the protein kinase p38 in the spindle assembly checkpoint and mitotic arrest. *Science* **280**, 599–602 (1998).

19. Widmann, C. *et al.* MEK kinase 1, a substrate for DEVD-directed caspases, is involved in genotoxin-induced apoptosis. *Mol. Cell. Biol.* **18**, 2416–2429 (1998).
20. Machida, Y. *et al.* Plant MAP kinase cascade that is mediated by MAPKKK-related kinase NPK1: Possible involvement in the regulation of the M phase (40th NIBB Conf. Stress response, 1998).
21. Nakashima, M. *et al.* The expression pattern of the gene for NPK1 protein kinase related to mitogen-activated protein kinase kinase (MAPKKK) in tobacco plant: correlation with cell proliferation. *Plant Cell Physiol.* **39**, 690–700 (1998).
22. Leyser, O. Auxin signalling: Protein stability as a versatile control target. *Curr. Biol.* **8**, R305–R307 (1998).
23. del Pozo, J. C. *et al.* The ubiquitin-related protein RUB1 and auxin response in arabidopsis. *Science* **280**, 1760–1763 (1998).
24. Guilfoyle, T. J. Aux/IAA proteins and auxin signal transduction. *Trends Plant Sci.* **3**, 205–207 (1998).
25. Ulmasov, T. *et al.* Aux/IAA proteins repress expression of reporter genes containing natural and highly active synthetic auxin response elements. *Plant Cell* **9**, 1963–1971 (1997).
26. Kyriakis, J. M. & Avruch, J. Sounding the alarm: protein kinase cascades activated by stress and inflammation. *J. Biol. Chem.* **271**, 24313–24316 (1996).
27. Nishihama, R. *et al.* Possible involvement of different splicing in regulation of the activity of *Arabidopsis* ANP1 that is related to mitogen-activated protein kinase kinases. *Plant J.* **12**, 39–48 (1997).
28. Klimczak, L. J. *et al.* Multiple isoforms of arabidopsis casein kinase I combine conserved catalytic domains with variable carboxyl-terminal extensions. *Plant Physiol.* **109**, 687–696 (1995).
29. Gleave, A. P. A versatile binary vector system with a T-DNA organizational structure conducive to efficient integration of cloned DNA into the plant genome. *Plant Mol. Biol.* **20**, 1203–1207 (1992).
30. Jang, J.-C. *et al.* Hexokinase as a sugar sensor in higher plants. *Plant Cell* **9**, 5–19 (1997).

Acknowledgements. We thank T. Ulmasov and T. J. Guilfoyle for the GH3-GUS construct; J.-C. Jang for the tobacco cDNA; M. Minet for the *Arabidopsis* cDNA library; H. Sun and N. K. Tonks for the MKP1 cDNA; and T. Jones, A. Khokhlatchev, G. Lazar, J. F. Nardone and J. Nishio for comments on the manuscript. J.S. is supported by Hoechst A.G., USDA and NSF.

Correspondence and requests for materials should be addressed to J.S. (e-mail: sheen@frodo.mgh.harvard.edu).

Ribonuclease E is a 5'-end-dependent endonuclease

George A. Mackie

Department of Biochemistry & Molecular Biology, University of British Columbia, 2146 Health Sciences Mall, Vancouver, British Columbia V6T 1Z3, Canada

The selective degradation of messenger RNAs enables cells to regulate the levels of particular mRNAs in response to changes in the environment. Ribonuclease (RNase) E (ref. 1), a single-strand-specific endonuclease^{2–4} that is found in a multi-enzyme complex known as the 'degradosome'^{5–7}, initiates the degradation of many mRNAs in *Escherichia coli*^{3,8,9}. Its relative lack of sequence specificity and the presence of many potential cleavage sites in mRNA substrates^{2,3} cannot explain why mRNA decay frequently proceeds in a net 5'-to-3' direction^{9–11}. I have prepared covalently closed circular derivatives of natural substrates, the *rpsT* mRNA encoding ribosomal protein S20 (ref. 2) and the 9S precursor to 5S ribosomal RNA^{1,12}, and find that these derivatives are considerably more resistant to cleavage *in vitro* by RNase E than are linear molecules. Moreover, antisense oligodeoxynucleotides complementary to the 5' end of linear substrates significantly reduce the latter's susceptibility to attack by RNase E. Finally, natural substrates with terminal 5'-triphosphate groups are poorly cleaved by RNase E *in vitro*, whereas 5' monophosphorylated substrates are strongly preferred (compare with ref. 13). These results show that RNase E has inherent vectorial properties, with its activity depending on the 5' end of its substrates; this can account for the direction of mRNA decay in *E. coli*, the phenomenon of 'all or none' mRNA decay, and the stabilization provided by 5' stem-loop structures^{14–17}.

If initiation of mRNA decay by RNase E begins only at or near the 5' end of an mRNA^{14–17}, then blocking or sequestering the 5' end of a substrate might spare it from attack. Circular RNAs lack a free 5' terminus; I prepared circular derivatives of *rpsT* mRNA by self-ligation, catalysed by DNA ligase¹⁸ in the presence of ATP and a 'bridging' oligodeoxynucleotide complementary to 12 residues at the 5' and 3' ends of the substrate. A 3' extension permitted efficient annealing (compare lanes 1 and 7 in Fig. 1a). Up to 20% of the input

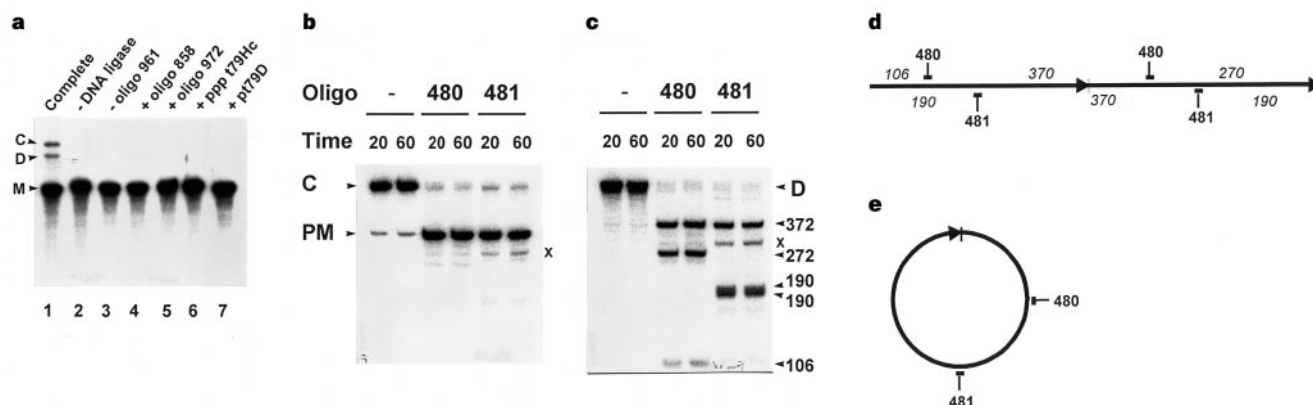


Figure 1 Preparation and characterization of covalently closed circular *rpsT* mRNA. **a**, RNA (~20 nM pt79Hc) and 300 nM oligonucleotide 961 were incubated with 1 U T4 DNA ligase (see Methods). Omissions (–) or substitutions (+) are indicated. Products include circles (C), linear dimers (D), and linear monomers (M). **b**, **c**, Oligonucleotide-directed RNase H digestion of circles (**b**) and dimers (**c**).

RNA was converted to forms with slower mobility, denoted by C or D in Fig. 1a, in a reaction dependent on DNA ligase (lane 2), a 5'-terminal monophosphate group (compare lanes 1 and 6), and oligonucleotide 961 (lane 3). Putative circular RNAs could be largely separated from linear monomers or dimers by preparative gel electrophoresis. Positive identification of circular *rpsT* mRNA (form C in Fig. 1a) was obtained by its conversion to a single species with mobility identical to the linear starting material (that is, a permuted monomer) by using oligonucleotide-directed cleavage with RNase H (Fig. 1b, e). In contrast, linear dimers (form D in Fig. 1a), a major byproduct of the ligation reaction, yielded three products whose sizes were dependent on the oligonucleotide (Fig. 1c, d).

The susceptibility of each form of *rpsT* mRNA to endonucleolytic cleavage was assayed using RNase E from two sources: highly purified degradosomes^{5,7,19} or essentially homogeneous, electrophoretically purified Rne protein²⁰. Assays also contained a standard substrate, 9S RNA, in 2–4-fold mass excess (Fig. 2a–d). The exonucleolytic activity of polynucleotide phosphorylase in the degradosome was suppressed by withholding phosphate, an essential cofactor. The rates of cleavage of linear and circular RNAs differed significantly, regardless of the source of RNase E activity. Circular *rpsT* mRNA (form C in Fig. 2) remained largely intact after 60 min of digestion with degradosomes (Fig. 2a) or with electrophoretically purified Rne protein (Fig. 2c). In contrast, the rate of formation of p5S RNA from 9S RNA (the internal control) by degradosomes was very rapid, with 50% of the input substrate disappearing within 2 min (Fig. 2a). The relative rate of cleavage of 9S RNA was slower with purified Rne protein, as observed previously²⁰, but was still at least 40-fold faster than cleavage of circular *rpsT* mRNA (Fig. 2c). To confirm that circular *rpsT* mRNA had not been inactivated during preparation, it was linearized by treatment with RNase H and oligonucleotide 961, then treated with RNase E in a second digestion (Fig. 2b, d). Between 95–100% of the circular *rpsT* mRNA was opened by the initial treatment, with the rest being opened during the second incubation (Fig. 2d). The opened circles were fully susceptible to digestion by RNase E, 50% disappearing within 1 min (Fig. 2b) or within 5 min (Fig. 2d). Discrete bands whose sizes were consistent with the position of known RNase E cleavage sites and the extent of RNase H cleavage were obtained with degradosomes (Fig. 2b; bands indicated by arrows). Primer extension confirmed that these bands were the result of endonucleolytic cleavages. The data in Fig. 2e show that the major site of cleavage of linearized circular *rpsT* mRNA occurs between residues 300 and 301, with another lesser site between residues 301 and 302 (Fig. 2e, lanes 1–3); these sites are identical to

those found in the positive control (lanes 4–6), in linear t79Hc RNA that had also been treated with RNase H and oligonucleotide 961 (lanes 7–9), and in *rpsT* mRNA *in vivo* and *in vitro*^{2,21}. The bands obtained with purified Rne were more heterogeneous (Fig. 2d), possibly because of nibbling of the 3' ends of the products by Rne itself²², but were otherwise similar to those obtained with degradosomes.

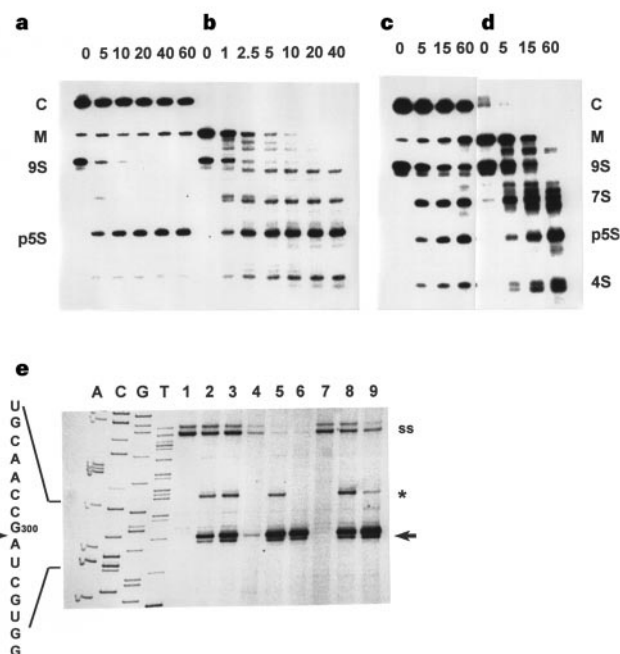


Figure 2 Cleavage of circular or linear pt79Hc. Circular pt79Hc (C) was incubated directly (**a**, **c**) or after treatment with 50 nM oligonucleotide 961 and 1 U RNase H for 20 min at 30°C (**b**, **d**) with purified degradosomes (**a**, **b**, **e**) or purified Rne protein (**c**, **d**) (see Methods). 20 nM 9S RNA (246 nt) was an internal standard; its product is p5S (126 nt). Positions of pt79Hc (M; ~390 nt) and its major RNase E product (160–165 nt) (arrow) are shown (see the text). **e**, RNAs digested for 0 (lanes 1, 4, 7), 5 (lanes 2, 5, 8) or 30 (lanes 3, 6, 9) min were used as templates for primer extension with oligonucleotide 15 (ref. 21): lanes 1–3, gel-purified circular pt79Hc opened with RNase H and oligonucleotide 961; lanes 4–6, pppt79D; lanes 7–9, gel-purified linear pt79Hc, also preincubated with RNase H and oligonucleotide 961. The mRNA sequence at the major cleavage site is shown beside the sequence ladder (primed with 5'-[³²P]oligonucleotide 15) and the corresponding cDNA by an arrow in the right margin. Symbols: ss, a doublet due to a strong stop; asterisk, a minor cleavage product at residues 291/292 (ref. 21).

To test the generality of the resistance of a circular RNA to RNase E, a second circular substrate was prepared from pt9SB RNA by self-ligation dependent on oligonucleotide 971. Circles (form C in Fig. 3a) were the predominant product (40% yield) and were distinguished from linear dimers (form D in Fig. 3a) by forming permuted linear monomers after oligonucleotide-directed digestion with RNase H (data not shown). As was the case for circular *rpsT* mRNA, circular 9S RNA was highly resistant to cleavage by RNase E in degradosomes (Fig. 3b) or by purified Rne protein (not shown) under conditions where linear 9S RNA was converted to its product, p5S RNA, by two cleavages. Linearization of the circular 9S RNA with RNase H and oligonucleotide 971 fully restored the ability of RNase E to cleave it to yield p5S RNA (Fig. 3c) at a rate virtually indistinguishable from that of linear molecules (data not shown). The mobility of the product obtained from opened circular molecules was identical to that from unmodified linear molecules, as assessed by migration on a sequencing gel (Fig. 3d: compare lanes 2, 6 and 7 with lanes 4 and 9–11).

To test whether an RNA needs to be covalently closed as a circle in order to acquire resistance to RNase E, oligodeoxynucleotides (Fig. 4a) were annealed to the *rpsT* mRNA (pt79D) and the heteroduplex digested with degradosomes or with Rne protein. Separate digestions with RNase H showed that the expected hybrids had formed (not shown). Cleavage of the control RNAs (pppt79D, pt79D, pppt79Hc and pt79Hc; Fig. 4c), as well as the heteroduplexes, obeyed first-order kinetics (data not shown). Oligonucleotide 858, complementary to the extreme 5' end of the RNA, reduced its rate of disappearance almost 5 to 12-fold, using degradosomes or Rne protein, respectively. Oligonucleotides 974, 975 or 976 each anneal to 13 successive residues in pt79D, leaving 2, 4 or 6 unpaired residues at the extreme 5' end (Fig. 4a). The data shown in Fig. 4b indicate that the initial rate of cleavage of such heteroduplexes by degradosomes increases with the number of unpaired 5' residues to ~70% of the control rate in the case of oligonucleotide 976. A similar trend is observed in the initial rates of cleavage by the

purified Rne protein. By contrast, cleavage of a heteroduplex with oligonucleotide 480, which leaves 94 unpaired 5' residues, proceeded at control rates (data not shown). In all cases, a normal spectrum of endonucleolytic products accumulated during digestion (not shown). None of the oligonucleotides used in Fig. 4b caused any reduction in the rate of cleavage of 9S RNA, even at 300 nM (data not shown). These and the previous data show that RNase E in the degradosome prefers an unpaired 5'-end in order that it may recognize any of the 12 characterized sites in the *rpsT* mRNA substrate², and that the inhibitory effect of a 5'-terminal stem-loop structure on the activity of RNase E *in vivo*¹⁵ is probably due to a direct effect on RNase E itself.

The data shown in Fig. 4c also indicate that the presence of a terminal 5'-triphosphate group strongly inhibits (20–30-fold) RNase E cleavage of 9S RNA and of two derivatives of *rpsT*

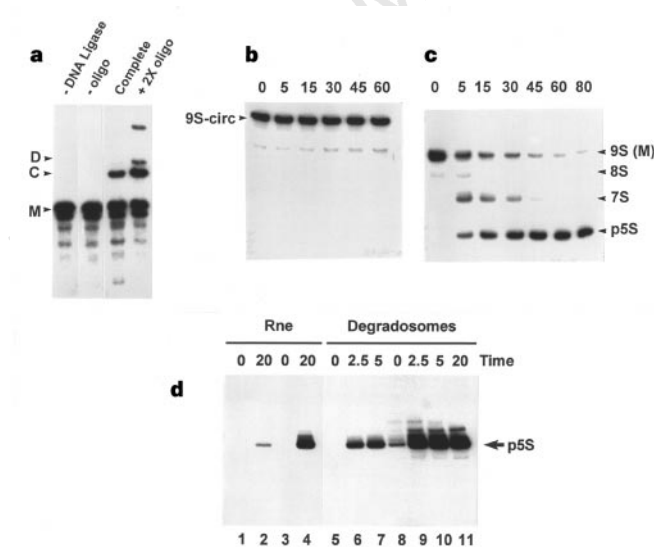


Figure 3 Circular 9S RNA. **a**, Circular 9S RNA was prepared from pt9SB (see Methods) using oligonucleotide 971 as the bridge. Omissions (–) are indicated above the first two lanes. The ligation in the fourth lane contained excess oligonucleotide 971. Positions of linear dimers (D), circles (C) and unligated monomers (M) are shown. **b**, **c**, Digestion of circular pt9SB RNA by purified degradosomes before (**b**) or after (**c**) digestion with RNase H and oligonucleotide 971. Positions of circular (C) or linear (M) pt9S RNA and of known products¹¹² 8S, 7S, and p5S are shown. **d**, Separation of products of the following substrates on a 6% sequencing gel: lanes 1, 2, 5, 6 and 7, circular 9S RNA (~5–10 nM) opened with oligonucleotide 971 and RNase H; lanes 3, 4, 8, 9 and 10, linear 9S RNA (≥20 nM); and lane 11, ppp9S RNA (20 nM). The 126-nt p5S product is indicated.

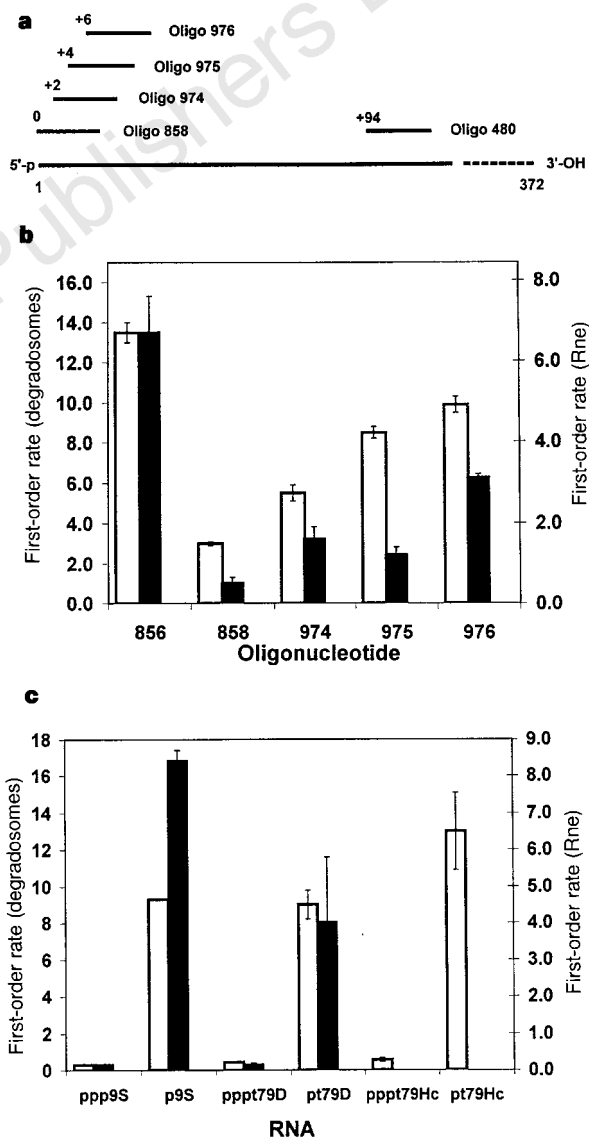


Figure 4 Cleavage of heteroduplex or differentially phosphorylated RNAs. **a**, Annealing of oligonucleotides of pt79D RNA. Residues 1–372 in the substrate correspond to residues 76–447 in previous reports^{2,7,21,27}. **b**, First-order rates of cleavage of pt79D in the presence of excess complementary oligonucleotides. Substrate (20 nM) was digested with limiting amounts of degradosomes (white bars) or purified Rne protein (black bars) (see Methods). Products were separated electrophoretically and the first-order rates calculated based on 2–5 determinations. **c**, Rates of cleavage of mono (p)- or tri (ppp)-phosphorylated RNAs. The substrates (20 nM) indicated along the x-axis were digested and products analysed as for **b**.

mRNA, t79D and t79Hc, compared to their monophosphorylated forms. Primer extension showed that the principal RNase E cleavage site in both forms of the *rpsT* mRNA occurred between residues 300–301, as expected (data not shown; see also Fig. 2e). These results extend the observation made previously that pppRNA_{L5} is degraded more slowly *in vivo* than pRNA_{L5} generated by RNase E cleavage^{13,23}, a result presumed to reflect the existence of a 5' to 3' exonuclease¹³, or a functional interaction between RNase E and polynucleotide phosphorylase²³. The data in Fig. 4 show that the protective effect of the terminal triphosphate is probably directed against RNase E itself, rather than to accessory proteins.

My results show that RNase E must recognize both an internal cleavage site and a 5'-terminal unpaired nucleoside monophosphate residue when it acts as an endoribonuclease. Whether a similar requirement exists for the recently reported exoribonuclease activity of Rne²² is unclear. To my knowledge, such a property is unprecedented among endoribonucleases. In addition to its catalytic site, RNase E should contain a phosphate-binding pocket to interact with the terminal nucleotide. The putative phosphate pocket must be sufficiently shallow to exclude triphosphates or polymeric RNA and sufficiently narrow to exclude base pairing to the 5' nucleotide. The location of such features in the Rne protein is unknown, but is amenable to experimental identification. Although kinetic parameters cannot be measured for each cleavage site in a complex substrate such as the *rpsT* mRNA, the initial rates of decay shown in Fig. 4c suggest that the affinity of RNase E for monophosphorylated (p) RNAs is significantly (20–30-fold) higher than for native (ppp) RNAs. Consequently, once degradation of an individual mRNA is initiated, continued digestion will be kinetically preferred over cleavage of intact pppRNAs, even if RNase E dissociates from its products after each successive cleavage. A key (but not initial) step in the degradation of eukaryotic mRNAs is the removal of the 5' cap^{24,25} to enable a 5' to 3' exonuclease, such as XRN1 in yeast, to act^{24,28}. The 5' triphosphate group in unprocessed transcripts in *E. coli* would functionally mimic the cap structure by constituting a rate-limiting barrier to RNase E, thereby protecting the body of an mRNA from attack. It is possible that cleavage of the α - β phosphoanhydride linkage at the 5' end of mRNAs could constitute the initiating step in mRNA decay, but such an activity is unknown in *E. coli*. □

Methods

Purification and assay of enzymes. Degradosomes were purified from strain CF881 lacking RNase I (ref. 26) using a modification¹⁹ of ref. 7 to a level of purity and activity equal to those of published methods^{7,8}. Electrophoretically purified Rne protein was obtained as described²⁰. RNase E activity was assayed² with 4.7 $\mu\text{g ml}^{-1}$ degradosomes or $\sim 1 \mu\text{g ml}^{-1}$ purified Rne protein and substrates at 1–20 nM, as noted in the figure legends. Pretreatment of RNAs with RNase H (1 unit) and a complementary oligonucleotide (60 nM)²⁷ and primer extension has been described^{21,27}.

Preparation of RNA. The DNA template pGM79 (ref. 21) encompassing the *rpsT* mRNA was linearized with either *DraI* or *HincII*, the latter to extend the transcript by 18 residues to permit annealing to oligonucleotide 961. Transcription with SP6 RNA polymerase in the presence of [α -³²P]CTP as described²¹ yielded t79D (372 nt) or t79Hc (390 nt). The template pJG9S-2, similar to pRC9S¹² but substituting an SP6 for the T7 promoter, was linearized with *BamHI* (to add a single-stranded 3' extension) and transcribed to yield t9SB (264 nt). Such transcripts are 5'-triphosphorylated (ppp). RNA was passed over a Sephacryl S-400 spin-column (Pharmacia) and digested with 0.2 units of calf intestinal alkaline phosphatase (Pharmacia) in the recommended buffer for 30 min at 37 °C. After extraction, this product was incubated with 0.5 mM ATP and 10 units T4 polynucleotide kinase (NEB) for 30 min at 37 °C to form monophosphorylated RNA (pRNA). Approximately 5–8 pmol of pRNA was heated in 30 mM Tris-HCl, pH 7.8, 10 mM MgCl₂, 10 mM DTT, for 2 min at 50 °C and 10 min at 37 °C with an empirically determined concentration of bridging oligonucleotide, then incubated for 5–15 h at 30 °C with 12 Weiss units of T4 DNA ligase (Pharmacia) and 0.5 mM ATP¹⁸. After

extraction and precipitation, the ligated RNAs were denatured by boiling in buffered 90% formamide and separated on a preparative 6% polyacrylamide gel containing 8M urea. RNAs were detected by phosphorimaging and recovered by electroelution.

Oligonucleotides. Oligonucleotide 15 (5'-AGCTTCAGCAAATTGGCG) is complementary to residues near the 3' terminus of *rpsT* mRNA²¹. Oligonucleotides 480 and 481 are complementary to residues 185–199 and 265–281, respectively, of t79D or t79Hc (ref. 27). Oligo 856 (5'-AGGCATGATGTGTTGC) was used as a non-complementary control; oligo 858 (5'-ATTCCGTG-TATTC) is complementary to the 5' ends of t79Hc and t79D; oligo 961 (5'-TTCCGTGATTCGACTCTAGAGGA) is complementary to the 5' and 3' termini of t79Hc; oligo 971 (5'-CCAAAACAGCTTCGATCCTCTAGAG) is complementary to the 5' and 3' termini of t9SB; oligo 972 (5'-GACTCTA-GAGGA) is complementary to the 3' terminus of t79Hc; oligos 974 (5'-GAATCCGTGTAT), 975 (5'-GGGAATCCGTGT) and 976 (5'-AGGG-GAATCCGT) are complementary to 13 residues in pt79D beginning 2, 4, or six residues from the 5' end (Fig. 4a).

Received 13 May; accepted 21 July 1998.

- Misra, T. K. & Apirion, D. RNase E, an RNA processing enzyme from *Escherichia coli*. *J. Biol. Chem.* **254**, 11154–11159 (1979).
- Mackie, G. A. Secondary structure of the mRNA for ribosomal protein S20. *J. Biol. Chem.* **267**, 1054–1061 (1992).
- Ehretmann, C. P., Carpousis, A. J. & Krusch, H. M. Specificity of *Escherichia coli* endoribonuclease RNase E: *in vivo* and *in vitro* analysis of mutants in a bacteriophage T4 mRNA processing site. *Genes Dev.* **6**, 149–159 (1992).
- McDowall, K. J., Lin-Chao, S. & Cohen, S. N. A+U content rather than a particular nucleotide order determines the specificity of RNase E cleavage. *J. Biol. Chem.* **269**, 10790–10796 (1994).
- Carpousis, A. J., Van Houwe, G., Ehretmann, C. & Krusch, H. Copurification of *E. coli* RNase E and PNase: evidence for a specific association between two enzymes important in mRNA processing and degradation. *Cell* **76**, 889–900 (1994).
- Miczak, A., Kabardin, V. R., Wei, C.-L. & Lin-Chao, S. Proteins associated with RNase E in a multicomponent ribonucleolytic complex. *Proc. Natl Acad. Sci. USA* **93**, 3865–3869 (1996).
- Py, B., Higgins, C. E., Krusch, H. & Carpousis, A. J. A DEAD-box RNA helicase in the *Escherichia coli* degradosome. *Nature* **381**, 169–172 (1996).
- Meleforts, Ö., Lundberg, U. & von Gabain, A. in *Control of Messenger RNA Stability* (eds Belasco, J. & Brawerman, G.), 53–70 (Academic, San Diego, 1993).
- Nierlich, D. P. & Murakawa, G. J. The decay of bacterial messenger RNA. *Progr. Nucleic Acid Res. Mol. Biol.* **52**, 153–216 (1996).
- Apirion, D. Degradation of RNA in *Escherichia coli*: a hypothesis. *Mol. Gen. Genet.* **122**, 313–322 (1972).
- Kennell, D. in *Maximising Gene Expression* (eds Reznikoff, W. S. & Gold, L.) 101–142 (Butterworths, Stoneham, MA, 1986).
- Cormack, R. S. & Mackie, G. A. Structural requirements for the processing of *Escherichia coli* 5S ribosomal RNA by RNase E *in vitro*. *J. Mol. Biol.* **228**, 1078–1090 (1992).
- Lin-Chao, S. & Cohen, S. N. The rate of processing and degradation of antisense RNA1 regulates the replication of ColE1-type plasmids *in vivo*. *Cell* **65**, 1233–1242 (1991).
- Emory, S. A., Bouvet, P. & Belasco, J. G. A 5' terminal stem-loop structure can stabilize mRNA in *Escherichia coli*. *Genes Dev.* **6**, 135–148 (1992).
- Bouvet, P. & Belasco, J. G. Control of RNase E-mediated RNA degradation by 5'-terminal base pairing in *E. coli*. *Nature* **360**, 488–491 (1992).
- Hansen, M. J. *et al.* The *ompA* 5' untranslated region impedes a major pathway for mRNA degradation in *Escherichia coli*. *Mol. Microbiol.* **12**, 707–716 (1994).
- Mackie, G. A., Genereaux, J. G. & Masterman, S. K. Modulation of the activity of RNase E *in vitro* by RNA sequences and secondary structures 5' to cleavage sites. *J. Biol. Chem.* **272**, 609–616 (1997).
- Moore, M. J. & Sharp, P. A. Site-specific modification of pre-mRNA: the 2'-hydroxyl groups at splice sites. *Science* **256**, 992–997 (1992).
- Coburn, G. A. & Mackie, G. A. Reconstitution of the degradation of the mRNA for ribosomal protein S20 with purified enzymes. *J. Mol. Biol.* **279**, 1061–1074 (1998).
- Cormack, R. S., Genereaux, J. G. & Mackie, G. A. RNase E activity is conferred by a single polypeptide: overexpression, purification, and properties of the *ams/rne/hmp1* gene product. *Proc. Natl Acad. Sci. USA* **90**, 9006–9010 (1993).
- Mackie, G. A. Specific endonucleolytic cleavage of the mRNA for ribosomal protein S20 of *Escherichia coli* requires the product of the *ams* gene *in vivo* and *in vitro*. *J. Bacteriol.* **173**, 2488–2497 (1991).
- Huang, H., Liao, J. & Cohen, S. N. Poly(A)- and poly(U)-specific RNA 3' tail shortening by *E. coli* ribonuclease E. *Nature* **391**, 99–102 (1998).
- Xu, F. & Cohen, S. N. RNA degradation in *Escherichia coli* regulated by 3' adenylation and 5' phosphorylation. *Nature* **374**, 180–183 (1995).
- Beelman, C. A. *et al.* An essential component of the decapping enzyme required for normal rates of mRNA turnover. *Nature* **382**, 642–646 (1996).
- Couttet, P. *et al.* Messenger RNA deadenylation precedes decapping in mammalian cells. *Proc. Natl Acad. Sci. USA* **94**, 5628–5633 (1997).
- Zhu, L., Gangopadhyay, T., Padmanabha, K. P. & Deutscher, M. P. *Escherichia coli* rna gene encoding RNase E: cloning, overexpression, subcellular distribution of the enzyme and use of an rna deletion to identify additional RNases. *J. Bacteriol.* **172**, 3146–3151 (1990).
- Mackie, G. A. & Genereaux, J. G. The role of RNA structure in determining RNase E-dependent cleavage sites in the mRNA for ribosomal protein S20 *in vitro*. *J. Mol. Biol.* **234**, 998–1012 (1993).
- Stevens, A. An exoribonuclease from *Saccharomyces cerevisiae*: effect of modifications of 5' end groups on the hydrolysis of substrates to 5' mononucleotides. *Biochem. Biophys. Res. Commun.* **81**, 656–661 (1978).

Acknowledgements. I thank G. Coburn for preparing degradosomes, A. Prud'homme-Genereux for helping to prepare Fig. 4, members of my laboratory for comments, and the MRC of Canada for an operating grant.

Correspondence and requests for materials should be addressed to the author (e-mail: gamackie@unixg.ubc.ca).

erratum

Subsurface charge accumulation imaging of a quantum Hall liquid

S. H. Tessmer, P. I. Glicofridis, R. C. Ashoori, L. S. Levitov & M. R. Melloch

Nature **392**, 51–54 (1998)

In this Letter in the 5 March issue, some of the superscript lettering on Figs 2–4 was printed incorrectly. In addition, the contrast was low in some copies of the journal. Reprints with the figures correctly produced are available, and they are reprinted here. □

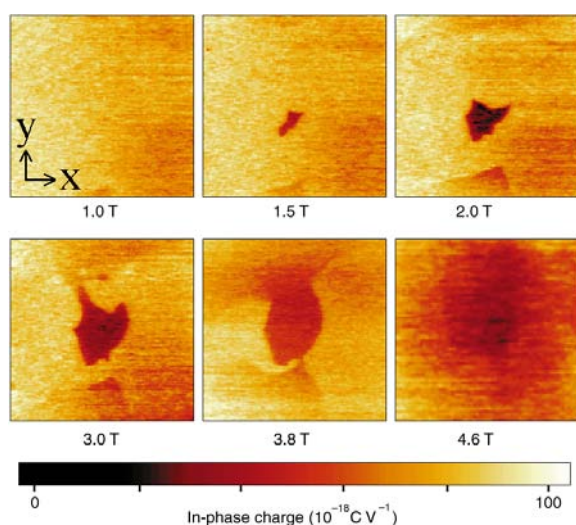


Figure 2

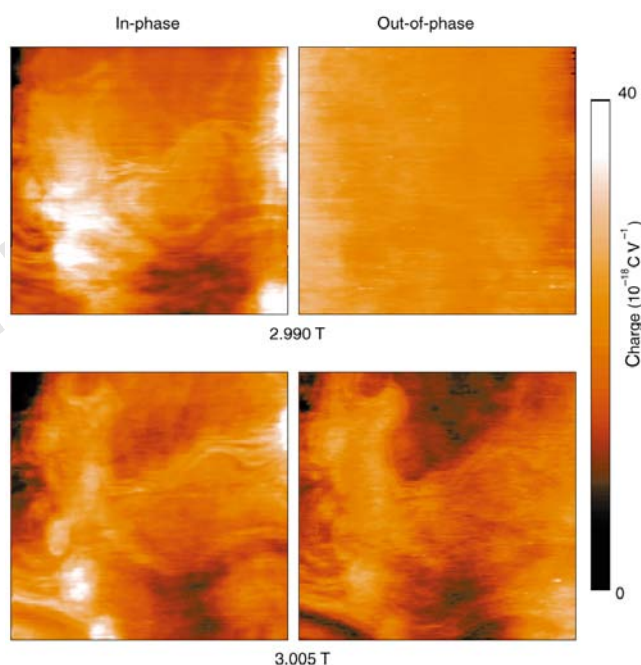


Figure 4

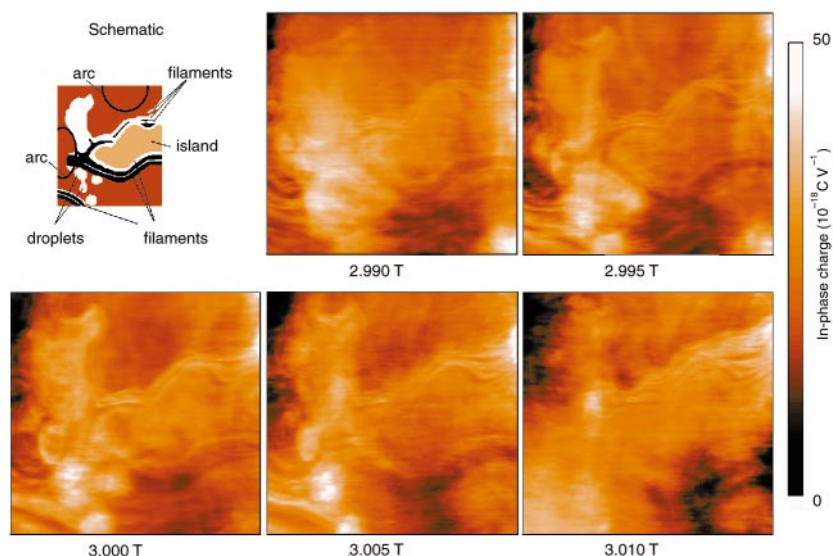


Figure 3

Ways to handle DNA

The molecular biology tools presented here include microchip array analysers and labels, telomeric probes, cloning kits for cosmic library construction and a system for quantifying and analysing fluorescently labelled nucleic acids.

E-Gels

From Invitrogen

Bufferless, precast agarose gels specifically designed for efficient electrophoresis

Comprising a complete electrophoresis system, these units are both gels and electrophoresis units. Agarose, electrodes and ethidium bromide are all packaged inside a dry, disposable, ultraviolet-transparent cassette. This is said to offer one-step electrophoresis and staining with excellent resolution and run times of only 35 minutes. E-Gels are supplied in a 12-well, minigel format with 1.2, 2.0 and 4.0 per cent agarose concentrations.

Reader Enquiry No. 100

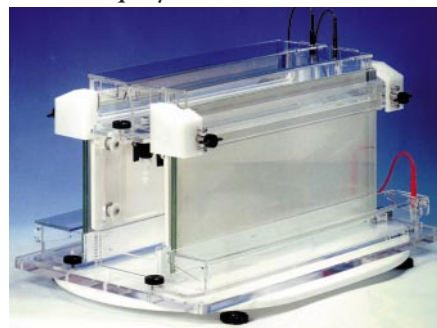
JumboGel system

From Hitachi Genetic Systems

A new addition to the company's line of fluorescence imaging products

Designed to maximize throughput for electrophoresis, this system facilitates loading and allows imaging of up to 264 samples in under three hours. The dual-sided gel electrophoresis apparatus can process two 45 × 28-cm (W × H) gels simultaneously, and uses a rotating base for easy access to both gels. For maximum sample throughput, the unit is 45-cm wide and accommodates up to 132 samples per gel. Sharktooth combs with 3-mm, point-to-point well spacing are used to accommodate multi-channel pipettors for efficient gel loading. Specially etched, low-fluorescence plates allow common pipette tips to fit accurately into the plate troughs, making alignment easier and eliminating the need for expensive gel-loading pipette tips. Other features include epoxy-coated, aluminium heat-dispersion plates to reduce gel 'smiling', bar clamps that evenly compress the glass plate sandwich to the upper reservoir gasket and permanent flip-lid covers with safety tip power leads.

Reader Enquiry No. 101



Big is as big does: JumboGel gel electrophoresis.



VP-ITC for measuring biomolecular interactions.

VP-ITC microcalorimeter

From MicroCal

A third generation, isothermal titration calorimeter (ITC) for the characterization of all types of biomolecular interactions

The ITC is designed to determine not only the binding constant and stoichiometry of an interaction, but also to estimate the heat and entropy of binding — all in about an hour. This system is said to be highly automated and to require no further operator attention once the sample cell and the injection syringe are filled with the macromolecule and ligand solutions, respectively. The instrument is controlled by VPViewer software with measurement parameters (such as temperature, number of injections and injection volume) being input by the operator. For minimum downtime between experiments that use differing temperatures, on-board thermoelectric cooling is used. Data analysis and graphical presentations are carried out with Origin software, utilizing stored mathematical routines.

Reader Enquiry No. 102

pWEB cosmid cloning kit

From Epicentre Technologies

An optimized kit for the production of cosmic libraries and for cosmic cloning

This kit eliminates the need for *Sau3A* partial digests through the use of sheared, end-repaired DNA. The requirement of intact genomic DNA of at least 200 kb for partial digestion is also eliminated, as it is not necessary to optimize the partial restriction-digest conditions. By simply shearing the DNA with a syringe, a truly random population of fragments to a clone is produced, which is not possible with *Sau3A*-generated fragments. The kit comes with a pre-digested, dephosphorylated pWEB vector; end-repair buffer and enzyme; FastLink DNA ligase; GELase preparation for purification of 30–45-kb fragments from LMP agarose gels; highly efficient MaxPlax packaging extracts; and T7 control DNA.

Reader Enquiry No. 103

Wave systems

From Transgenomic

Automated, high-throughput mutation scanning and polymorphisms detection

With this DNA fragment analysis system and DNasep technology, mutations appear on the system software as a characteristic pattern of peaks. These peaks correspond to the mixture of DNA heteroduplexes and homoduplexes formed when wild-type and mutant DNA are hybridized. Four key aspects of mutation detection can be analysed: PCR primer design, PCR protocol, separation gradient and separation temperature.

Reader Enquiry No. 104

Microarrays, probes and analysis

GeneTAC biochip analyser

From Genomic Solutions

A compact, bench-top unit that automates the imaging and analysis of gene microarrays

This system consists of an imager and an analysis workstation. The imager uses a high-resolution CCD camera in a light-tight enclosure. It can detect spots of various diameters quickly and accurately from low- to medium-density arrays that have been labelled with CY3 or CY5. Customization is available for detection of other fluorescent dyes, and the sample holder allows the user to load up to 24 samples at a time for 'walk-away' operation. The workstation includes high-density array analysis software for rapid identification, quantification and reporting of spots.

Reader Enquiry No. 105

BioProbe

From Enzo Diagnostics

Nucleic acid labelling systems formulated specifically for use in microchip array assays

Systems are offered for labelling RNA or DNA with either biotin or fluorescein. Each kit includes all labelling reagents, a complete protocol and biotin-labelled or fluorescein-labelled nucleotides. Enzo offers three labelling kits, optimized for use in microchip array assays. Amplified and fragmented DNA can be efficiently labelled at the 3'-hydroxy

terminus with either biotin-modified or fluorescein-modified dideoxynucleotides, using the 3'-hydroxy terminal labelling system for microchip arrays. The BioProbe RNA transcript system enables production of fluorescein-labelled RNA using either T3 or T7 polymerases; whereas, the high-yield transcript labelling system utilizes T7 RNA polymerase and both biotinylated UTP and biotinylated CTP for highly efficient labelling of RNA in yields of 100 µg or greater.

Reader Enquiry No. 106

Lumi-Imager F1

From Roche Diagnostics

A workstation that can visualize, quantify and analyse fluorescently labelled nucleic acids, proteins and chemiluminescent probes

The system uses a high-performance, cooled CCD camera to capture images with high sensitivity and resolution, and in short exposure times. The need for X-ray film or multiple film exposures is eliminated, and fluorescent labels can be visualized directly in the gel. The system is flexible enough to accommodate an extensive range of DNA, RNA, protein and microtitre plate analyses. It uses a focus-stabilized lens optimized to detect low-light signals. The unit's fluorescence excitation source is a software-controlled, UV transilluminator that has been incorporated into the sample stage. A five-position filter wheel gives three standard fluorescent detection options (520, 600 and 645 nm), together with the option of additional filters for other fluorescent dyes. Running under Windows NT, the Pentium PC supplied with the system is configured with an integral magnetic optical drive for high-capacity, removable data storage and archiving.

Reader Enquiry No. 107

TelVysion DNA Probes

From Vysis

DNA probes for the detection of genetic abnormalities involving telomeres

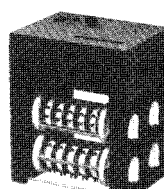
Intended for prenatal, congenital and cancer research, these probes utilize fluorescence *in situ* hybridization to detect visually cryptic, hard-to-identify translocations. For research use only, these probes specifically identify the telomeric regions of DNA, detecting rearrangements and deletions that occur at the ends of chromosomes, and that cannot be identified using standard cytogenetic analysis. The probes are directly labelled with fluorescent molecules, providing visual signals that are easy to interpret and make the detection of these chromosome anomalies unambiguous.

Reader Enquiry No. 108 □

These notes are compiled by Brendan Horton from information provided by the manufacturers. For more details, fill in the reader service card bound inside the journal.

ADVERTISEMENTS

DIANDERM®
EQUILIBRIUM DIALYSIS SYSTEM



The gentlest method to quantify the binding parameters and thermodynamic data between ligands and biopolymers. Results come closest to the actual values. Sample volume from 0.2 ml to 5.0 ml. Up to 20 experiments in parallel.

DIANDERM
Gentle
P.O. Box 65012
D 81215 Munich

READER ENQUIRY NO. 61

PIERCE
GelCode® Blue Stain Reagent
Eliminates the Need for Methanol/Acetic Acid Destaining

Features/Benefits:

- Works without the smelly methanol/acetic acid destaining step that is common with traditional stains
- Provides substantially improved sensitivity by way of a Water Wash Enhancement™ step
- Ready-to-use formula provides one-step, one-hour staining
- Bands develop before your eyes and can be viewed directly in the staining tray
- Crystal-clear background with no shrinkage
- No increased background with overnight staining and no loss of sensitivity from over-destaining

Tel: 800-874-3723 or 815-968-0747 • Fax: 815-968-7316
CS@piercenet.com • Internet: www.piercenet.com

READER ENQUIRY NO. 48

NEWCASTLE UNIVERSITY
MOLECULAR BIOLOGY UNIT

PROTEIN SEQUENCING
Membrane-bound, solid or solution samples
£12 per cycle
5 cycles minimum - no set up charge

For more information please contact Dr. Joe Gray:
Phone/Fax: +44 191 222 8612
Email: JOE.GRAY@newcastle.ac.uk

DNA sequencing (ABI 377-96XL, 877 Catalyst) also available at competitive prices

READER ENQUIRY NO. 42

Wako BioProducts
Wako Chemicals USA, Inc.
1600 Bellwood Road
Richmond, VA 23237
Telephone: (804) 271-7677
Facsimile: (804) 271-7791
(800) 992-9256
bioproducts@wakousa.com

Wako

Detect residual DNA with Wako's DNA Extractor Kit
or Isolate Genomic DNA with DNA Extractor WB Kit
Both use the Nal Method
No Phenol No Chloroform

READER ENQUIRY NO. 39

SAINT
THE TRANSFECTION INNOVATORS

Easy
•
Efficient
•
Economic
+

<http://www.saintnl.com>
• Info@saintnl.com
fax +31 50 3118995

READER ENQUIRY NO. 49

Severn Biotech Ltd
Customer Services
Comprehensive Custom Service for DNA, Peptides and Polyclonal antibodies from SBL

- DNA Synthesis
- Multi Peptide Synthesis (5mgs)
- Larger Scale Peptide Synthesis 25 - 100mgs
- HPLC Purification/Mass Spectroscopy
- Custom Conjugation to Carrier Proteins
- Custom Polyclonal Antibodies

For a Fast Efficient Service Reply Today
Severn Biotech Ltd, Unit 2, Park Lane, Stourport Road, Kidderminster, Worcs DY11 6TJ, UK
Tel: (01562) 825286 Fax: (01562) 825284
E-mail: 101626.2102@compuserve.com